

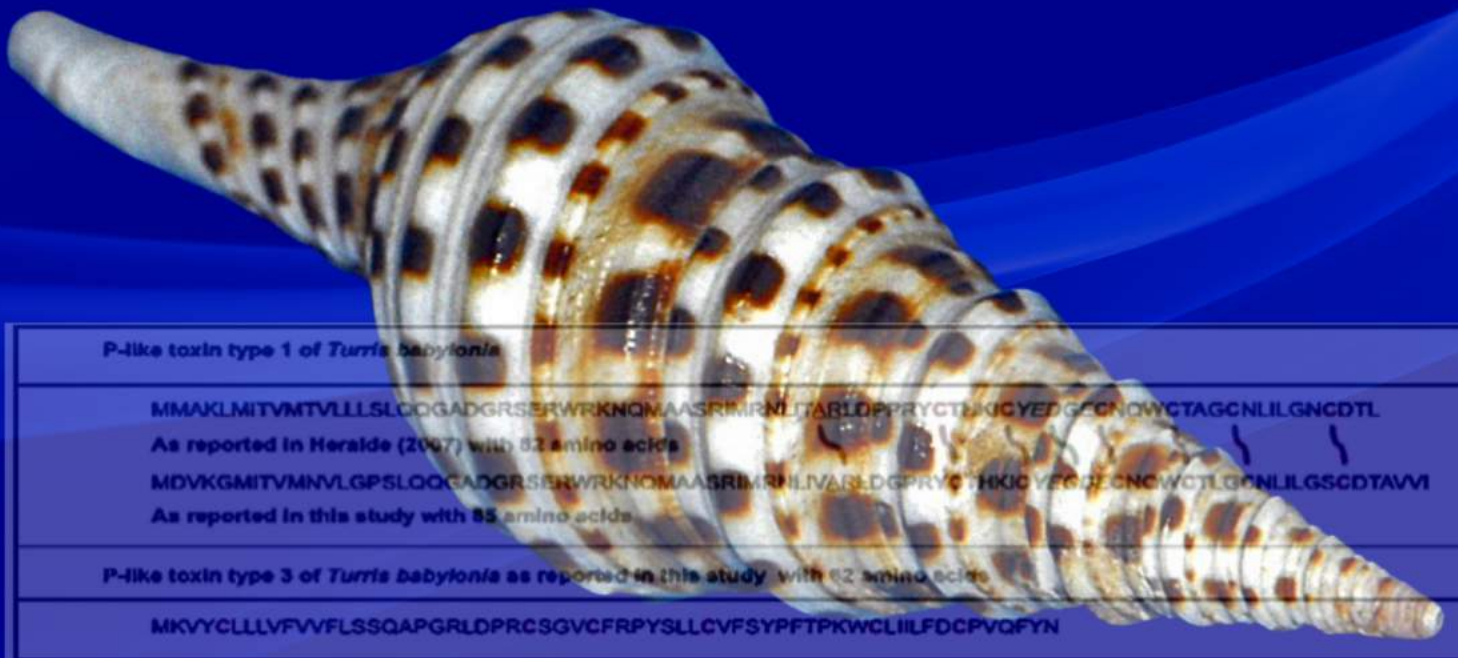
# PJBMB

The Philippine Journal of  
Biochemistry and Molecular Biology  
(formerly Bulletin of the Philippine Biochemical Society)

JANUARY-DECEMBER 2020 | VOLUME 1 | ISSUE 1&2 | ISSN: 2719-1990

Published by: The Philippine Society of Biochemistry and Molecular Biology

Published online at [www.psbmb.org/pjbmb](http://www.psbmb.org/pjbmb)



P-like toxin type 1 of *Turris babylonis*

MMAKLMITVMTVLLLSLQGGADGRSERWRKNQMAASRIMRNLTARLDPPRYCTDKICYEDGECNQCWCTAGCNLILGNCDTL

As reported in *Heralde* (2007) with 83 amino acids

MDVKGMITVMNVLGPSLQGGADGRSERWRKNQMAASRIMRNLTARLDPPRYCTDKICYEDGECNQCWCTAGCNLILGSCDTAVVI

As reported in this study with 85 amino acids

P-like toxin type 3 of *Turris babylonis* as reported in this study with 62 amino acids

MKYYCLLLVFVFLSSQAPGRLDPRCSGVCFRPYSLLCVFSYPFTPKWCLLILFDCPVQFYN



January-December 2020 | Volume 1, Issue 1&2

### *The Editorial Board*

#### **Editor-in-Chief:**

Francisco M. Heralde III, RN. PhD.

#### **Members:**

Leslie Michelle M. Dalmacio, PhD  
Mudjekeewis D. Santos, PhD  
Crist John M. Pastor, PhD  
Ian Kendrick C. Fontanilla, PhD  
Mary Ann O. Torio, PhD  
Marcos B. Valdez, Jr., PhD  
Neil Andrew D. Bascos, PhD  
Arlen A. Dela Cruz, PhD  
Roberta N. Garcia, PhD  
Ma. Cristina Francesca T. Dimaculangan, PhD  
Orlex B. Yllano, PhD

#### **Reviewers for this Issue:**

Crist John Pastor, *PNU (Philippines)*  
Elmer Mojica, *Pace University (USA)*  
Gracia Fe B. Yu, *UP Manila (Philippines)*  
Ian Kendrick C. Fontanilla, *UP Diliman (Philippines)*  
Jose Isagani B. Janairo, *DLSU (Philippines)*  
Jun Guevarra, *UST (Philippines)*  
Leslie Michelle M. Dalmacio, *UP Manila (Philippines)*  
Marcos B. Valdez Jr., *FEU (Philippines)*  
Maritess D. Cation, *UST (Philippines)*  
Mudjekeewis D. Santos, *NFRDI (Philippines)*  
Muhammad Naeem, *BZU (Pakistan)*  
Nathaniel Hepowit, *Vanderbilt University (USA)*  
M. Prassad Naidu, *NMCH-Nellore (India)*  
Roberta N. Garcia, *UPLB (Philippines)*

The Philippine Journal of Biochemistry and Molecular Biology (PJBMB) (formerly Bulletin of the Philippine Biochemical Society) is a peer-reviewed journal published biannually by the Philippine Society of Biochemistry and Molecular Biology (PSBMB) which covers original research articles and research communications in biochemistry, molecular biology and/or other related disciplines, biochemical education articles, biochemical literature reviews, special interest articles with biochemical applications, how-to articles on useful activities relating to biochemistry, letters and book reviews.

All communications and inquiries can be addressed to:

**Francisco M. Heralde III, RN. PhD.**  
**Email: [pjbmb.editor.2020@gmail.com](mailto:pjbmb.editor.2020@gmail.com)**

#### **About the Cover**

An image of a *Turris babylonica* from the article by Hilario et. al. "P-like conotoxins detected in *Turris babylonica*". Photo credits to Hectonichus.



January-December 2020 | Volume 1, Issue 1&2

## Table of Contents

<b>Message from the Incumbent PSBMB President</b> .....	i
<i>Leslie Michelle M. Dalmacio</i>	
<b>Editorial</b> .....	ii
<i>Francisco M. Heralde III</i>	
<b>DNA Barcoding of Confiscated Endangered Philippine Sailfin Lizard <i>Hydrosaurus pustulatus</i> (Eschsholtz, 1829) (Short Communication)</b> .....	1
<i>Rogel Victor D. Mendoza and Ian Kendrick C. Fontanilla</i>	
<b>P-like conotoxins detected in <i>Turris babylonia</i> (Short Communication)</b> .....	7
<i>Allan L. Hilario and Francisco M. Heralde III</i>	
<b>Mass Spectrometry and Proteomics as Emerging Technologies for Breast Cancer (Review)</b> .....	12
<i>Maritess D. Cation and Maria Cristina Ramos</i>	
<b>Identification of Class I HLA Alleles in Anonymized Cell Therapy Specimens through Real-Time PCR with Melt-Curve Analysis (Short Communication)</b> .....	29
<i>Joanne Jennifer E. Tan, Maria Teresa A. Barzaga and Francisco M. Heralde III</i>	
<b><i>In Silico</i> Pathway Analysis of the Anti-cancer Mechanism of Selected Active Components of Virgin Coconut Oil and their Key Targets (Full Article)</b> .....	37
<i>Excellces Dee Montemayor, Mayrell Ann F. Ravina, Jay T. Dalet, Francisco M. Heralde</i>	
<b>Preliminary Evaluation of rbcL, matK, and SRAP Markers for the Molecular Characterization of Five Philippine <i>Allium sativum</i> varieties (Short Communication)</b> .....	54
<i>Allisandra Isabel B. Bigtas, Patrick Gabriel G. Moreno, and Francisco III M. Heralde III</i>	
<b>Preliminary characterization and <i>in silico</i> studies on the alpha-amylase inhibitors from <i>Momordica charantia</i> AMP06 methanolic leaf extract (Short Communication)</b> .....	61
<i>Arra B. Asejo<sup>1</sup>, Patrick G. Moreno, Junie Billones, Ruel Nacario, and Francisco M. Heralde III</i>	
<b>Guide to Contributors, Manuscript Preparation, Categories of Articles</b> .....	71



January-December 2020 | Volume 1, Issue 1&2

## *Message from the Incumbent PSBMB President*

The Philippine Society for Biochemistry and Molecular Biology (PSBMB) pursues the promotion of scientific fields covered by the society through conferences and seminars. It is therefore with joy and pride that the *Philippine Journal of Biochemistry and Molecular Biology (PJBMB)*, which features biochemistry and molecular biology research in the Philippines, is launched at this time when the value of sound research cannot be overemphasized.

The *PJBMB* is formerly *the Bulletin of the Philippine Biochemical Society*, initiated by the pillars of Philippine biochemistry and molecular biology. The current PSBMB Board of Directors hope that the *PJBMB* will be another venue for the research outputs of its members and those in the field of biochemistry and molecular biology in the academe, industry and research institutions. The *PJBMB* will also serve as a rich source of new knowledge and inspiration for future research.

For greater reach, *PJBMB* is published biannually as an open access journal. It will feature original research articles and research communications in biochemistry, molecular biology and/or other related disciplines, biochemical education articles, biochemical literature reviews, special interest articles with biochemical applications, how-to articles on useful activities relating to biochemistry, letters and book reviews.

We enjoin the Philippine biochemistry and molecular biology community to support the journal through submission of your latest research output. Through the efforts of the current Editor-in-Chief and the Editorial Board, may *PJBMB* thrive and succeed for the benefit of the nation.

Leslie Michelle M. Dalmacio, PhD.  
PSBMB President (2020-2021)

## **Editorial:**

### **On the Maiden Issue, Toxins, Pandemic and the Challenges Ahead!**

The last two years has been a blend of mix experiences. In 2018, when we launch the Philippine Journal of Biochemistry and Molecular Biology (PJBMB) at the PSBMB National Convention in Iloilo City; there were lots of things we do not know like what will happen in the coming years. What was clear then in our mind was to usher the maiden issue of PJBMB as an online publication to cater to the needs of the growing number of PSBMB members. We hold the idea of reviving and continuing the tradition started under the Philippine Biochemical Society (i.e., the former and predecessor of the present PSBMB) of producing and maintaining a Society Research Journal, the Bulletin of the Philippine Biochemical Society.

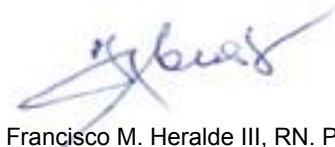
The idea of producing the maiden issue has been a challenge from the time that the articles are solicited, to the time the articles are reviewed and later on layed-out into the Journal we want it to be. Thus, the maiden issue has not come without its own share of angst and birthing pains, but as the saying goes, when its time to be debuted into the world comes, nothing can take away that feeling of joy, fulfillment, and success. As Editor-In-Chief, I am sharing this feeling to all the members of the PSBMB.

Seven articles are featured in this maiden issue covering various topics from the molecular biology of fishes, gastropods and plants to the biochemistry and diagnostics of cancer and immune markers and natural products in silico studies and enzyme inhibition. Indeed, the topics have a wide array of coverage and reflects on the diversity of local researches on the aspects of biochemistry and molecular biology.

The choice of *Turris babylonia* as a feature cover of this maiden issue stems from the long tradition of peptide toxin research that spans from the pioneer founding members of the Society like the National Scientist Dr. Lourdes Cruz and the Harvard Distinguished Professor Dr. Baldomero Olivera who started the conotoxin research in the Department of Biochemistry and Molecular Biology, UP College of Medicine, continuing in Marine Science Institute in UP Diliman with Dr. Gisela Conception and moving to the University of Utah and globally maturing into a full-blown field of study in gastropod toxinology, becoming a springboard for several new generations of PhDs in the field of biochemistry and molecular biology.

In 2019 and 2020, we saw a big contrast in research activities both locally and globally, among different institutions including the PSBMB. The CoVID19 pandemic has literally reduced research activities in various aspects and have mostly focused on health pertinent to the SARS-CoV2. The nature of the researches have also transformed to largely in silico and clinical, while some with daring laboratories are able to pursue their research agenda with limited funding in various academic institutions which were all beset with lockdowns, social distancing and other restriction to contain the deadly viral spread. Of course, this impacted on the kind of researches that ended up in publications. Indeed, the pandemic has shaped and impacted on the lives of everyone and has continued to change the way things operate like our classes that were mostly online, without face-to-face contact, minimal University-based laboratories opened, and Society Conferences being migrated to virtual formats.

A lot of challenges lie ahead for everyone to maintain and continue to propagate our kind of science. The question and issue of the vaccines and the different types available commercially presented a scenario for various stakeholders to look back and engage in an active discussion to thresh-out the safety issues, the urgency and pertinence to the wider population. Among the scientist and researchers, the discussion may proceed further into the biochemistry and molecular biology principles concerning vaccine development, the relevance of the platform used and the emergence of mutant strains and as to how they impact on vaccine efficacy. While the pandemic remains and continues to ravage the population, and without a protective guarantee from the vaccine in eliciting herd immunity, the challenge remains and this defines the new normal where our kind of science will exist, remain, and move forward.



Francisco M. Heralde III, RN. PhD.  
Editor-In-Chief

# DNA Barcoding of Confiscated Endangered Philippine Sailfin Lizard *Hydrosaurus pustulatus* (Eschsholtz, 1829)

Rogel Victor D. Mendoza<sup>1,\*</sup>, Ian Kendrick C. Fontanilla<sup>1,2</sup>

<sup>1</sup> DNA Barcoding Laboratory, Institute of Biology, University of the Philippines, Diliman, Quezon City 1101, Philippines

<sup>2</sup> Philippine Genome Center, University of the Philippines, Diliman, Quezon City 1101, Philippines)

\* Corresponding author

## Email address:

rdmendoza1@up.edu.ph

## To cite:

Mendoza, RVD; Fontanilla, IKC.2020. DNA Barcoding of Confiscated Endangered Philippine Sailfin Lizard *Hydrosaurus pustulatus* (Eschsholtz, 1829). PJBMB. Vol. I, No. 1, 2020, pp. 01-06. doi: 10.5555/pjbmb.ph.2020.01.01.01

Received: 08 24, 2019; Accepted: 09 25, 2020; Published: 12 02, 2020

**Abstract:** The Philippines, a known biodiversity hotspot, faces the threat of higher extinction rates due to disruptive human activities that include illegal wildlife trade. Among those species affected by these activities is the endemic Philippine Sailfin Lizard (*Hydrosaurus pustulatus*), which is currently listed as 'Vulnerable' by the International Conservation of Natural Resources (IUCN). Proper identification of illegally traded samples could be addressed using DNA barcoding that uses small segments of DNA in confirming species identity. *Cox1* gene has been used in DNA barcoding *H. pustulatus*. However, the 16S rRNA gene region, which is also commonly used as a marker for reptiles, is lacking in the species. Here, we demonstrated the use of DNA barcodes, specifically *cytochrome oxidase subunit I (cox1)* genes, in the confirmation of identity of illegally traded *H. pustulatus* and its possible geographic origin. We also provided novel 16S gene sequences together with *cox1* gene to infer its placement within the Agamidae. BLAST results show that the generated *cox1* barcodes matched with *H. pustulatus* (100%) found in GenBank. Furthermore, haplotype network analysis revealed that the confiscated samples were similar to the *H. pustulatus* haplotype found in Polillo Island. Phylogenetic analysis using concatenated 16S rRNA and *cox1* genes showed that *H. pustulatus* clustered with a congeneric, *H. amboinensis* (ML Bootstrap=100; NJ Bootstrap=100). In this study, we demonstrated that DNA barcodes could aid not only in the proper identification of the species but also their possible geographic origin. This could be useful in providing data on hotspot areas of wildlife trafficking. In addition, the use of the 16S gene can potentially be used together with *cox1* in discriminating between *H. pustulatus* and *H. amboinensis*.

**Keywords:** DNA Barcoding, Wildlife Forensics, Philippine Sailfin Lizard, *cox1*, 16S

## 1. INTRODUCTION

The Philippines is considered as one of the biodiversity hotspots in the world with its high levels of endemism coupled with higher risk of extinction due to human activities such as energy use and land conversion which leads to habitat loss (Ehrlich 1994; Cincotta et al. 2000; Myers et al. 2000). In addition, wildlife trafficking has been a major challenge in Asia with millions of animals being illegally exported in the region alone (Nijman 2010).

Among those species facing illegal wildlife trade is the endemic Philippine sailfin lizard, *Hydrosaurus pustulatus* (Eschsholtz 1829), which is categorized as 'Vulnerable' by the International Conservation of Natural

Resources (IUCN) (Ledesma 2009). This omnivorous reptile is found in several islands of the Philippines including Polillo, Mindoro, Negros, Guimaras, Panay, Masbate, Tablas, Romblon, Sibuyan, Catanduanes, Bicol, and other small isolated islands where its identification is confused with its congeneric species, *H. amboinensis* (Ledesma 2009; Siler et al. 2011).

*H. pustulatus* is a squamate reptile classified under subfamily Hydrosaurinae, along with its congeneric *H. amboinensis*, of family Agamidae (Pyron et al. 2013). Agamidae is a monophyletic family considered as the Old World counterpart of the New World Iguanidae and is a sister clade to Chamaelonidae based on their 16S and 12S mitochondrial rRNA genes (Honda et al. 2000). However, *H. pustulatus* is not represented in this

phylogenetic analysis due to lack of available molecular data.

DNA barcoding could be utilized in the rapid identification of known and novel species by using combinations of nucleotides found in DNA to produce unique barcodes that could discriminate different species (Hebert et al. 2003; von Crautlein et al. 2011). The use of mitochondrial genes has been demonstrated to allow the identification of cryptic species especially among animals (Hebert et al. 2003; Feng et al. 2011; Luo et al. 2011). However, DNA barcoding efforts in the Philippines has not been significantly documented, with some taxa lacking representation in the genetic database (Fontanilla et al. 2014). In addition, reptiles lack efficient universal primers that target the commonly used animal mitochondrial gene, the cytochrome oxidase subunit I (*cox1*) (Vences et al. 2012). One solution is the use of the 16S rRNA gene, which currently serves as the suitable mitochondrial DNA marker for the taxon (Vences et al. 2012). However, no 16S rRNA gene has been published in the genetic database for the identification of *H. pustulatus*.

In this study, we aimed to utilize the *cox1* gene to confirm the identity of confiscated *H. pustulatus* specimens and determine their possible geographic origin. Using sequences from the *cox1* and 16S genes, the placement of *H. pustulatus* within the Agamidae is elucidated.

## 2. METHODOLOGY

### 2.1 Acquisition of Samples

Seven vials of *H. pustulatus* tail clippings were used as the source of DNA. These came from samples that were confiscated by the Department of Environment and Natural Resources-Biodiversity Management Bureau (DENR-BMB) Wildlife Rescue Center (WRC) from an illegal consignment on 28 January 2016 at the Philippine Airlines (PAL) cargo terminal. This consignment was to be exported to Japan along with other confiscated endangered and endemic species including tarsiers, watersnakes, ratsnakes, and the Philippine scops owl.

### 2.2 DNA Extraction and Amplification

Tissue samples from the tail clippings were subjected to DNA extraction using the GeneJET Genomic DNA Purification Kit (Thermo Scientific) following the manufacturer's protocol.

Extracted DNA was subjected to PCR amplification using VF1: 5'TTCTCAACCAACCACAAAGACATTGG-3' (Ivanova et al. 2006) and VR1: 5'TAGACTTCTGGGTGGCCAAAGAATCA-3' (Ward et al. 2005) as the forward and reverse primers, respectively, in order to amplify the *cox1* gene. PCR was utilized using 1.0 µL of 10x PCR buffer, 36.25 µL distilled water, 0.25 µL (5units/ µL) Taq polymerase, 1.5 µL (1mM) forward and reverse primers, 2.5 µL (0.2mM)

DMSO, 2.0 µL of (50mM) MgCl<sub>2</sub>, and 5.0 µL (3-40 ng/µL) DNA template from the sample. The PCR thermal regime for the *cox1* gene consisted of an initial denaturation at 94°C for 2 mins, followed by 5 cycles of denaturation at 94°C for 40 secs, annealing at 45°C for 40 secs, and extension at 72°C for 1 min; the annealing temperature was then changed to 51°C and the thermal cycling ran for another 35 cycles before the final extension at 72°C for 5 cycles.

The 16S rRNA was amplified using 16S Ar: 5'-CGCCTGTTTATCAAAAACAT 3' and 16S Br: 5'-CCGGTCTGAACTCAGATCACGT-3' (Palumbi 1996) as the forward and reverse primers, respectively. The same PCR components followed the concentrations used in amplification of *cox1* except those of the DNA template and distilled water were adjusted to 1.0 µL and 40.25 µL, respectively. The PCR regime consisted of an initial denaturation at 96°C for 5 mins, followed by 43 cycles of denaturation at 96°C for 30 secs, annealing at 45°C for 30 secs, extension at 65°C for 1 min, and a final extension at 72°C for 5 mins.

### 2.3 Agarose Gel Electrophoresis and Sequencing

A 1% Agarose gel was prepared by dissolving one gram of agarose powder (Vivantis) in 100mL of 0.5X tris borate EDTA(TBE) buffer. One µL of 10mg/ml ethidium bromide (Invitrogen™) was added to the mixture for gel visualization. The gel mixture was allowed to solidify in a cassette and was submerged in 0.5X TBE buffer inside the AGE setup.

Each PCR product was loaded into the wells using 2 µL of Blue Juice as the loading dye. One µL of KAPPA Express DNA Ladder (KAPABiosystems) was used as the molecular weight marker. The setup was run for 30 minutes at 100 volts and was visualized under a UV illuminator after the run. Bands indicating the presence of the gene were cut given that they corresponded to their expected sizes (*cox1*=800bp; 16S=500bp). These bands were subjected to gel extraction and PCR clean-up using the QIAquick® Gel Extraction Kit (QIAGEN®), after which they were sent to 1st Base in Singapore for Sequencing.

### 2.4 Molecular Identification of confiscated *H. pustulatus* specimens

The consensus sequences for *cox1* and 16S genes were determined from generated forward and reverse reads using STADEN package (Staden et al. 2000). The consensus sequences were subjected to Basic Local Alignment Search Tool (BLAST) to identify the samples' closest match.

### 2.5 Determination of geographic location

To determine the possible geographic location of the confiscated samples, *H. pustulatus cox1* sequences were also downloaded from GenBank. These sequences represent different clades of *H. pustulatus*

haplotypes with known location based on the data published by Siler et al. (2014). Sequences were aligned using BioEdit v7.2.5 (Hall 2005) via the ClustalW (Thompson et al. 2003) function. Aligned sequences were trimmed using the program GBlocks v.0.91b (Castresana 2002). A median joining network was constructed using the R program packages pegas (Paradis 2010) and ape (Paradis 2018).

## 2.6 Determination of the placement of *H. pustulatus* within the Agamidae

Gene sequences of 16S and *cox1* from the whole mitochondrial genome of representatives of family Agamidae were downloaded in GenBank. These gene regions were aligned with the 16S and *cox1* gene sequences obtained from the confiscated *H. pustulatus* samples using the program BioEdit (Hall 2005). The aligned datasets were trimmed using the program GBlocks (Castresana 2002) for the sequences to be suitable for gene tree construction. Trimmed *cox1* and 16S gene sequences were concatenated using the program DAMBE (Xia & Xie 2001).

The model of substitution for the concatenated dataset for the ML tree construction was determined using jModelTest v2.1.10 (Guindon & Gascuel 2003; Durrin et al. 2012). Construction of maximum likelihood tree was carried out using IQTree (Nguyen et al. 2004). Bootstrap support of 1000 replicates were generated for neighbor-joining (Saitou & Nei 1987) and maximum likelihood (Felsenstein 1981) methods using PAUP 4.0 (Swofford 2002) and IQTree (Nguyen et al. 2004), respectively.

## 3. RESULTS AND DISCUSSION

### 3.1 Molecular confirmation of confiscated *H. pustulatus* specimens

All sequences generated from this study were submitted to the Barcode of Life Data (BOLD) with the following accessions: MN228919-MN228925 (*cox1*) and MN322560-MN322566 (16S). BLAST results using the *cox1* gene region showed that the confiscated specimens were identified as *H. pustulatus* with 100% identity. Its congeneric, *H. amboinensis*, was only 95.41%-96.18% identical with the samples (Table 1). Since there are currently no available 16S gene region sequences, BLAST results of the 16S gene region from the confiscated samples were most similar (99.1%) with its congeneric, *H. amboinensis* (Table 1). Furthermore, the haplotype network (Figure 1) based on the *cox1* gene of the confiscated *H. pustulatus* specimens suggest that the samples were identical to *H. pustulatus* found in Polillo Island and most similar to Aurora and suggests the possible origin of these samples from these areas.

*H. pustulatus* is one of the endemic species in the Philippines frequently targeted for illegal wildlife trade

Table 1. Basic local alignment search tool (BLAST) results for confiscated samples of *Hydrosaurus pustulatus*.

Voucher Specimen	Accession Number	Gene	BLAST Identified organism	% Identity	Accession Number
1A	MN228925	<i>cox1</i>	<i>Hydrosaurus pustulatus</i>	100.0%	KTO75334
	MN322566	16S	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
1B	MN228923	<i>cox1</i>	<i>Hydrosaurus pustulatus</i>	100.0%	KTO75334
	MN322564	16S	<i>Hydrosaurus cf. amboinensis</i>	96.1%	FJ952249
1C	MN228922	<i>cox1</i>	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
	MN322563	16S	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
2B	MN228924	<i>cox1</i>	<i>Hydrosaurus pustulatus</i>	100.0%	KTO75334
	MN322565	16S	<i>Hydrosaurus cf. amboinensis</i>	96.0%	FJ952249
4A	MN228921	<i>cox1</i>	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
	MN322562	16S	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
4B	MN228920	<i>cox1</i>	<i>Hydrosaurus pustulatus</i>	100.0%	KTO75334
	MN322561	16S	<i>Hydrosaurus amboinensis</i>	95.5%	AB475096
4C	MN228919	<i>cox1</i>	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096
	MN322560	16S	<i>Hydrosaurus amboinensis</i>	99.1%	AB475096

due to their striking dorsal crests, conspicuous caudal sail-like structure and ornate coloration (Siler et al. 2014). In fact, pet traders use different media including online sites in advertising the sale of Philippine Sailfin Lizards as pets. Fortunately, with the advent of DNA barcoding effort in the Philippines, identification of illegally traded wildlife became more effective and efficient. For example, in the Philippines, the use of *cox1* gene was proven to be useful in confirming the identity of some frozen dressed pangolins, which were confiscated from a Chinese fishing vessel in Tubbataha, to be that of the critically endangered Sunda pangolin *Manis javanica* (Luczon et al. 2016). This shows that the use of DNA barcoding could be useful when morphological identification becomes impossible due to circumstances such as sample decomposition or alteration.



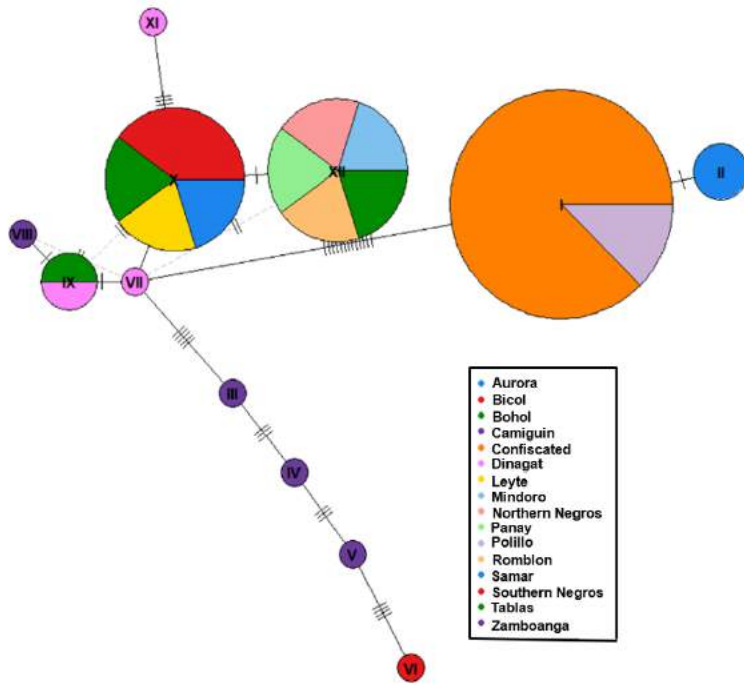


Figure 1. Haplotype Network of *H. pustulatus* haplotypes of the *cox1* gene from the confiscated samples and GenBank sequences used in the study. Location of the GenBank sequences are based on data published by Siler et al. (2014). Each bar represents a single nucleotide change.

### 3.2 Phylogenetic analysis using *cox1* and 16S rRNA gene sequences

Concatenation of the 16S rRNA and *cox1* genes resulted to a 796-nucleotide long sequence alignment which was used to infer the phylogenetic relationship in Agamidae using the Transversional model (TVM) (Posada 2003) with invariant sites (+I) and rate variation among sites (+G) as determined by jModelTest as the optimal model. The resulting ML tree supports the monophyly of Agamidae (ML Bootstrap=100; NJ Bootstrap=100). Furthermore, *H. pustulatus* clustered with *H. amboinensis*, supporting the grouping of the two congeneric species into subfamily Hydrosaurinae (ML Bootstrap=100; NJ Bootstrap=100). This tree also supports the clustering of subfamilies Agaminae (ML Bootstrap=99; NJ Bootstrap=99), Amphibolurinae (ML Bootstrap=100; NJ Bootstrap=97), Draconinae (ML Bootstrap=100; NJ Bootstrap=100), and Leiopinae (ML Bootstrap=100; NJ Bootstrap=100) of Agamidae (Figure 2).

Although no 16S rRNA gene region is present in the current genetic database for *H. pustulatus*, we were able to demonstrate the use of the gene, along with the *cox1* gene, in elucidating the relationships within the family. These genes could likewise distinguish *H. pustulatus* and *H. amboinensis* as observed in a distinct divergence between the two species (Figure 2). The inclusion of 16S rRNA gene region could also be useful in elucidating a wider phylogenetic study of the species with other reptiles due to the lack of *cox1* for the said group.

## 4. CONCLUSION

DNA barcodes are shown to be useful in confirming the identity of *H. pustulatus* as demonstrated by the *cox1* gene. Furthermore, the gene was also able to narrow down the likely origin of the trafficked *H. pustulatus* samples based on the *cox1* barcode data from Siler et al. (2014). In addition, we have also providing novel 16S gene sequences for the species, adding a gene marker for possible identification of the species. Nevertheless, creating a more extensive database that includes all possible localities where the species is located could provide a more comprehensive detection system of wildlife trafficking hotspots in the Philippines provided that there is sufficient variation to distinguish populations from across the different localities. Determination of the most likely areas where illegal wildlife trafficking is rampant could be useful from the perspective of law enforcement.

Furthermore, inclusion of the 16S gene was able to aid in inferring the relationship of *H. pustulatus* with other agamids by complementing the *cox1* database for the taxon. The concatenated dataset was also able to demonstrate the high support for various subfamilies within the Agamidae.

## ACKNOWLEDGMENT

We would like to thank the Department of Environment and Natural Resources Biodiversity Management Bureau (DENR-BMB) and BMB-Wildlife Rescue Center (WRC) for providing the samples used in this study and for the funding.

## REFERENCES

- Castresana J. 2002. GBLOCKS: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Version 0.91 b. Copyrighted by J. Castresana, EMBL.
- Cincotta RP, Wisnewski J, Engelman R. 2000. Human population in the biodiversity hotspots. *Nature*. 404 (6781): 990
- Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*. 9(8): 772

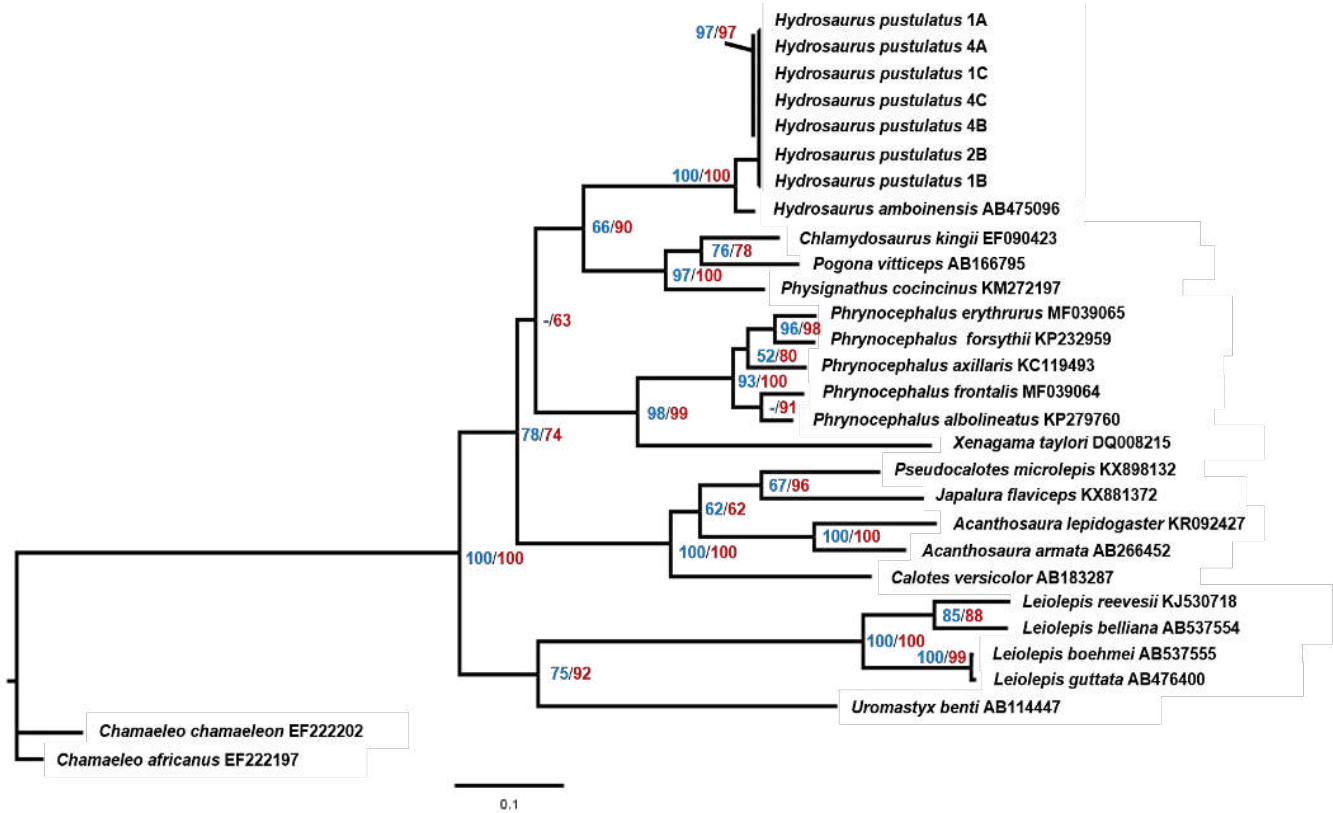


Figure 2. Maximum likelihood tree of Agamidae based on 796 nucleotides of the concatenated cox1 and 16S rRNA genes, with Chamaeleonidae (*Chamaeleo chamaeleon*, *Chamaeleo africanus*) as the outgroup. Values on the nodes represent percent bootstrap values based on 1000 bootstrap replicates using maximum likelihood (red) and neighbor-joining (blue) methods; values less than 50% are not shown. Scale bar represents one nucleotide substitution for every 10 nucleotides.

Ehrlich PR. 1994. Energy use and biodiversity loss. *Philos. Trans. Royal Soc. A.* 344(1307): 99-104.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17(6): 368-376.

Feng Y, Li Q, Kong L. 2011. DNA barcoding and phylogenetic analysis of Pectinidae (Mollusca: Bivalvia) based on mitochondrial COI and 16S rRNA genes. *Mol. Biol. Rep.* 38:291-299.

Fontanilla I, Torres A, Cañasa J, Yap S, Ong P. 2014. State of animal DNA barcoding in the Philippines: A review of COI sequencing of Philippine native fauna. *Philipp. Sci. Lett.* 7(1):104-137.

Guindon, S, Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52(5): 696-704.

Hall T. 2005. BioEdit: biological sequence alignment editor for Win95. *BioEdit. Biological sequence alignment editor for Win95.*

Hebert P, Cywinska A, Ball S, deWaard J. 2003. Biological identifications through DNA barcodes. *Proc R Soc Lond [Biol].* 270:313-321.

Honda M, Ota H, Kobayashi M, Nabhitabhata J, Yong H, Sengoku S, Hikida T. 2000. Phylogenetic Relationships of the Family Agamidae (Reptilia:Iguania) Inferred from Mitochondrial DNA Sequences. *Zool Sci.* 17:527-537.

Ivanova NV, Dewaard JR, Hebert PDN. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol. Ecol Notes.* 6(4): 998-1002.

Ledesma M, Brown R, Sy E, Rico EL. 2009. *Hydrosaurus pustulatus*. The IUCN Red List of Threatened Species 2009: e.T10335A3194587.

Luczon AU, Ong PS, Quilang JP, Fontanilla IKC. 2016. Determining species identity from confiscated pangolin remains using DNA barcoding. *Mitochondrial DNA B.* 1(1): 763-766.

Luo A, Zhang A, Ho S, Xu W, Zhang Y, Shi W, Cameron S, Zhu C. 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics.* 12:84.

Myers N, Mittermeier RA., Mittermeier CG, Da Fonseca GA, & Kent J. 2000. Biodiversity hotspots for conservation priorities. *Nature.* 403(6772): 853.

- Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32(1): 268-274.
- Nijman, V. 2010. An overview of international wildlife trade from Southeast Asia. *Biodivers Conserv.* 19 (4): 1101-1114.
- Palumbi SR. 1996. Nucleic acids II: The polymerase chain reaction. In: Hillis DM, Moritz C, Mable BK (eds) *Molecular Systematics*. Sinauer Associates, Inc. pp. 205-247.
- Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics.* 26(3): 419-420.
- Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics.* 35(3): 526-528.
- Posada D. 2003. Using MODELTEST and PAUP\* to select a model of nucleotide substitution. *Curr. Protoc. Bioinformatics.* (1): 6-5.
- Pyron R, Burbrink F, Wiens J. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol Biol.* 13:93.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4(4): 406-425.
- Siler C, Lira A, Brown R. 2014. Conservation genetics of Australasian sailfin lizards: Flagship species threatened by coastal development and insufficient protected area coverage. *Biol. Conserv.* 169:100-108.
- Siler C, Welton L, Siler J, Brown J, Bucol A, Diesmos A, Brown R. 2011. Amphibians and reptiles, Luzon Island, Aurora Province and Aurora Memorial National Park, Northern Philippines: new island distribution records. *Check List.* 7(2):182-195.
- Staden R, Beal KF, Bonfield JK. 2000. The staden package, 1998. In *Bioinformatics methods and protocols* (pp. 115-130). Humana Press, Totowa, NJ.
- Swofford DL. 2002. PAUP\* 4.0 b10. Phylogenetic analysis using parsimony (and other methods), version, 4, b10.
- Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512-526
- Thompson JD, Gibson TJ, Higgins DG. 2003. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics.* (1): 2-3.
- Vences M, Nagy Z, Sonet G, Verheyen E. 2012. DNA barcoding amphibians and reptiles. *DNA Barcodes.* 858:79-107.
- von Cräutlein M, Korpelainen H, Pietiläinen M, Rikkinen J. 2011. DNA barcoding: a tool for improved taxon identification and detection of species diversity. *Biodivers. Conserv.* 20(2): 373-389.
- Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN. 2005. DNA barcoding Australia's fish species. *Philos Trans R Soc Lond B Biol Sci.* 360(1462): 1847-1857.
- Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. *J. Hered.* 92(4): 371-373.

## P-like conotoxins detected in *Turris babylonia*

Allan L. Hilario<sup>1,\*</sup>, Francisco M. Heralde III<sup>1</sup>

<sup>1</sup> Department of Biochemistry and Molecular Biology, College of Medicine, University of the Philippines Manila

\* Corresponding author

### Email address:

alhilario@up.edu.ph

### To cite:

Hilario, AL; Heralde, FIIM.2020. P-like conotoxins detected in *Turris babylonia*. PJBMB. Vol. 1, No. 1, 2020, pp. 07-11.

10.5555/pjbmb.ph.2020.01.01.07

Received: 08 07, 2019; Accepted: 09 25, 2020; Published: 12 02, 2020

**Abstract:** Toxin homologs are common across turrid species and reflect on similarities of their prey types. P-like conotoxin is one type isolated in various turrid snails, which exhibits conserved sequences in their precursor peptides with marked diversity and subtype variations in similar groups of turrid snails, a toxin repertoire patterning, termed as “P-coding” system, employed by turrid snails as a toxin diversification strategy to better target their prey. This study aimed to determine if P-like type 1 and type 3 conotoxin genes are present in *Turris babylonia* using gene-specific primers used in *Gemmula* species. Total RNA was extracted from venom duct and then used to prepare double-stranded (ds) cDNA. The ds cDNA was used as template for P-like type 1 and type 3 conotoxins amplification. Then, the amplicons generated were sent to Macrogen, Inc., Seoul, South Korea for sequencing and analyzed using DNAsis. Results showed that P-like conotoxin type 1 and type 3 were amplified from *T. babylonia*. The gene sequences showed similar framework IX scaffold. The P-like conotoxin type 1 has 85 amino acid residues with the characteristic six-cysteine residues and a conserve region of YEDGE similar to *T. babylonia*. The P-like conotoxin type 3 is the first of this type ever reported in this gastropod species. It has 62 amino acid residues with six-cysteine residues but with divergent amino acid sequence from the P-like type 1 conotoxin. P-like conotoxin type 1 and 3 were detected in *T. babylonia* using gene specific primers for *Gemmula* species. The detection of these P-like conotoxins provides support on the hypothesis of a possible “P-coding” system among the turrid snails. Similar approach can be done in other turrid species.

**Keywords:** *Turris babylonia*, P-like conotoxins, venom peptides, conidae, turriptide

### 1. INTRODUCTION

Venomous marine snails of the superfamily Conoidea are considered one of the most diverse invertebrate lineages attributing to more than 10,000 species (Olivera *et al.* 2014; Imperial *et al.* 2014). Among the Conoidea, turrids are the most primitive, appearing around 120 million years ago (Heller 2015).

Turrids are a paraphyletic group from the marine Gastropoda and were previously designated as the family Turridae, encompassing more than 3,600 named living species with several new species described every year (Kantor *et al.* 2017). However, there is no distinct turrid shell shape by which all members can easily be identified (Omega *et al.* 2017). Due to numerous examples of homoplasy in their shell characteristics, there have been species delimitation and difficulty in

morphological identification of turrids (Kantor *et al.* 2017).

Nevertheless, there are few iconic turrid species, such as *Turris babylonia*, the “tower of babel” (see Figure 1) which has a characteristic shell shape, sculpture and color pattern, thus allowing an unambiguous identification. For 200 years, *Turris babylonia* has been considered an easily identifiable and nominate species for its genus (Kantor *et al.* 2017). In order to capture their prey, these snails are equipped with feeding guilds and complex venoms known as turritoxins which diverge between species through hypermutation within gene families, each having a specific target prey. Although they are still characterized with highly conserved signal and pro sequences (Olivera *et al.* 2014; Omega *et al.* 2017; Heller 2015).

Since the first characterization of turritoxins, numerous toxins are reported in the literature. Toxin homologs are common across turrid species and reflect on similarities of their prey-types. P-like conotoxin is one type isolated in various turrid snails. P-like conotoxins exhibit conserved sequences in their precursor peptides with marked diversity and subtype variation in similar groups of turrid snails, a toxin repertoire patterning, termed by Heralde (2007) as “P-coding” system, employed by turrid snails as a toxin diversification strategy to better target their prey. Hence, this study was conducted to

determine if P-like type 1 and type 3 conotoxin genes are present in *Turris babylonica* using gene-specific primers used in *Gemmula* species, a genus belonging to same family as *Turris babylonica*. The use of genus-specific primers among different genera in a family of marine snails like the turrids may prove the presence of conserved genes among the genus within the family and a simpler approach in conotoxin discovery

### 2.1. Sample Collection, Authentication, and Total RNA Extraction

The sample was taken from the authenticated collection of Francisco M. Heralde III and collected from the coasts of Cebu and Bohol, Philippines 300 to 500 meters off the coastline at the depth of 20 meters using tangled nets. The venom ducts were isolated, collected, and placed in RNeasy<sup>™</sup> (Ambion, TX, USA) and stored at -80°C until further use. Total RNA were first extracted from the venom ducts of the snails (*Turris babylonica*) with RNeasy<sup>™</sup> Mini Tissue Kit (Qiagen, Hilden, Germany) and kept at -80°C until further use.

### 2.2. Double stranded cDNA Preparation

Extracted total RNA from the venom ducts of the snails was then used in the preparation of the double-stranded (ds) cDNA, following the SMART cDNA (Clontech Laboratories, Inc., CA, USA) protocol following the manufacturer’s protocol.

### 2.3. PCR Amplification using primers for *Gemmula* Species

The putative conotoxin genes were amplified from two µL dsDNA product as template in 20 µL PCR reaction mix containing two µL PCR buffer (10X), dNTPs (200 µM), primers (0.5 µM each), Taq DNA polymerase (1U) and diethylpyrocarbonate (DEPC)-treated water. The primers used to amplify P-like conotoxin 1 was (5'-A(G/T)CGAAG(A/C)GCT(C/G)CATTCG-3') and P-like conotoxin 3 was (5'-ATC (G/C)A(T/G)(C/T)GAT (C/A)TGTT(T/G)TG-3') which were gene-specific primers designed for peptide conotoxins of *Gemmula speciosa*. The PCR profile used were as follows: an initial denaturation of one min at 95 °C, followed by 40 cycles



Figure 1. *Turris babylonica*. The shell with angular whorls. Taken from the coast of Cebu and Bohol, Philippines, May 2006, 20- meter depth, 6 cm. (Heralde, 2007) Reproduced with permission.

of 20 sec at 95 °C, 20 sec at 54 °C and 30 sec at 72 °C, and a final extension of 5 min at 72 °C. All amplicons were analyzed by gel electrophoresis using agarose gel.

### 2.4 DNA Sequencing

The cDNA was then used as template for P-like type 1 and type 3 conotoxin amplification. The amplicons were analyzed and visualized using gel electrophoresis and the generated amplicons were sent to Macrogen, Inc., Korea for sequencing. The sequence was analyzed for open reading frame and protein sequence determined using DNAsis (Miraibio, Inc., CA, USA).

## 3. RESULTS AND DISCUSSION

Turrids produce venom that interferes with the neuromuscular ion channels by preventing the prey from closing the sodium gates and opening the potassium gates, thereby disrupting the electric signals that leave the nerve cell. This eventually results to the continuous paralyzing twitch of all the body muscles, which then immobilizes their prey (Heller 2015; Gonzales & Saloma 2014). The turrid snails pose significant challenges in its taxonomy as being the largest family in the superfamily Conoidea and with largely unknown turrid peptides found in their venom (Olivera, Seron, & Fedosov 2010; Kendel *et al.* 2013).

The enormous resource of natural peptide toxins from this venom has great pharmacological and research potential. Each species has its own distinct complement of highly structured peptide toxins, which then has a specific, physiologically relevant protein target. The molecular, physiological and pharmacological characterizations of these diverse and numerous peptide toxins have evolved in a new class of bioactive drugs. These highly bio-diversified groups of venomous marine gastropods still have so much to be discovered for potential new species and pharmacologic uses. The peptide toxin discovery from conoideans has revolutionized the way scientists try to harvest the natural resources available for pharmacological application (Olivera *et al.* 2014; Heller 2015; Gonzales & Saloma 2014).

As applied in this study, the determination of P-like conotoxins using gene-specific primers developed from various turrids such as *Gemmula* species was successful in detecting similar toxins in *Turris babylonica*. In the study by Heralde (2008), the Pg-gene superfamily of *Gemmula* venom peptides are identified and characterized. The similar primers from *Gemmula speciosa* and *Gemmula lisajoni* coding for P-like conotoxin type 1 and type 3 respectively were used in this study.

The PCR amplification of the synthesized double stranded cDNA using the gene-specific primers for P-like conotoxin types 1 and 3 of *Gemmula* species showed distinct bands as seen from the agarose gel (see Figure 2). These bands corresponded to the expected amplicon size of approximately 250 bases. The utility of primers for *Gemmula* species in amplification of P-like conotoxin types 1 and 3 in *Turris babylonica* as used in this study provided evidence that the conotoxins

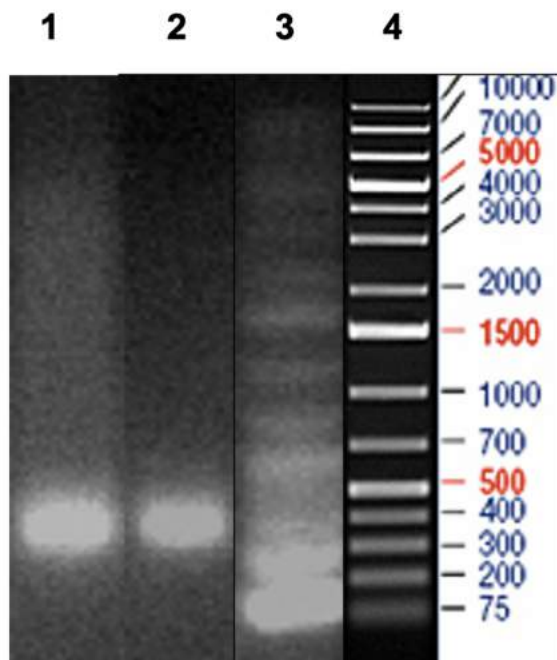


Figure 2. The agarose gel of PCR amplicons of P-like toxin type 1 and 3. Lane 1 is the P-like toxin type 1 PCR product. Lane 2 is the P-like toxin type 3. Lane 3 is the actual 1 kb GeneLadder™ run with the sample. Lane 4 is the 1 kb GeneLadder™ bands for comparison taken from the manufacturer's website (<http://www.fementas.com/>).

are conserved throughout evolution making this diverse group of marine gastropods as one of the most successful species of venomous marine organisms which use toxins for prey capture, avoidance of predation, and successful competition in the highly competitive marine ecosystem (Watkins, Hillyard, & Olivera 2006, Fu *et al.* 2018; Jin *et al.* 2019).

The gene sequences of P-like conotoxin types 1 and 3 showed similar framework IX scaffold (C-C-C-C-C-C)

with differences in the mature peptide sequences. The toxin precursor of P-like conotoxin type 1 has 85 amino acid residues with the characteristic six-cysteine residues and a conserve region of YEDGE similar to the one reported by Heralde (2007), also in *T. babylonica* (see Table 1). The toxin precursor of P-like conotoxin type 3 is the first of this type ever reported in this gastropod species. It has 62 amino acid residues with the same six-cysteine residues but with divergent amino acid sequence from the P-like type 1 conotoxin. The toxin precursor peptide sequences of both toxins are shown in Table 1. Despite the presence of structural similarities of P-like conotoxin 1 and 3 with the general conotoxins with the distinctive cysteine framework, the biological function of these toxins may not necessarily be similar. The idea that the structural characteristics of these toxins follows the purported biological functions may not hold true to this group of marine gastropods. There is such a wide functional diversity among conotoxins due to post-translational modifications, sequence variation of the mature peptide toxins, and structural framework variations (Miles *et al.* 2002).

The conotoxins are complex polypeptides composed of three major sequences wherein the polypeptides are cleaved to release the functionally mature peptide sequence. These are short sequence polypeptides with a signal sequence in the "pre" region at the N-terminal, followed by a pro sequence, and the biologically active sequence at the C-terminal part. The several superfamilies of conotoxins derived from the cone snails (*Conus*) are well studied than its more abundant cousins, the auger snails (*Hastula*) and the turrid snails (*Turrid*). The diversity of conotoxins found in cone snails is also expected among auger and turrid snails. Hence, the body of knowledge about the conotoxins provides the foundation in our effort to elucidate the functional diversity of peptide toxins in auger and turrid snails (Watkins, Hillyard, & Olivera 2006; Becker & Terlau 2008).

The detection of P-like conotoxins type 1 and type 3 in *Turris babylonica* strengthens the hypothesis of the possible "P-coding" system among the generic members of the subfamily *Turrinae* (H. Adams and A. Adams 1853 (1838)) and possibly among the *Conoidea*. Such P-like toxin gene homolog is present in most members of the subfamily *Turrinae*. In 2007, Heralde reported a toxin repertoire patterning, termed as "P-coding" system, which is used by these snails to effectively capture their prey. This coding system allows identification of potential toxins in Conoidean snails other than cone snails that have similar P-like pattern in its precursor peptide as used in categorizing conotoxins among cone snails (Heralde 2007; Heralde 2008; Robinson & Norton 2014). The approached used in this study to isolate potential peptide toxins in less studied and researched families of Conoidean snails can be

Table 1. Toxin precursor sequences of P-like conotoxin types 1 and 3 of *Turris babylonica* from amplicons using the designed P-like conotoxin type 1 and 3 primers for *Gemmula spp.* Peptide cleavage site is shown underlined while the conserved cysteine scaffold in bold letters. Conserved sequence relevant to clustering is shown italicized.

<b>P-like toxin type 1 of <i>Turris babylonica</i></b>	
MMAKLMITVMTVLLLLSLQQGADGRSERWRKNQMAASRIMRNLI <b>AR</b> LDPPRY <b>CTHKICYEDGECNQWCTAGCNLILGNCDTL</b>	
<b>As reported in Heralde (2007) with 82 amino acids</b>	
MDVKGMITVMNVLGPSLQQGADGRSERWRKNQMAASRIMRNLI <b>AB</b> LDGPRY <b>CTHKICYEGDECNQWCTLG</b> CNLI <b>LGSCDTAVVI</b>	
<b>As reported in this study with 85 amino acids</b>	
<b>P-like toxin type 3 of <i>Turris babylonica</i> as reported in this study with 62 amino acids</b>	
MKVY <b>C</b> LLL <b>V</b> FV <b>V</b> LLSSQAPGR <b>L</b> DP <b>R</b> C <b>S</b> GV <b>C</b> FR <b>P</b> YS <b>L</b> L <b>C</b> V <b>F</b> S <b>Y</b> P <b>F</b> TP <b>K</b> W <b>C</b> L <b>I</b> L <b>F</b> D <b>C</b> P <b>V</b> Q <b>F</b> Y <b>N</b>	

used to further the research initiatives and our understanding of the diversity of the turrid peptides.

#### 4. CONCLUSION

P-like conotoxin types 1 and 3 were detected in *T. babylonica* using gene specific primers for *Gemmula* species. The P-like type 3 toxin detected in this study was the first ever reported in this gastropod species. The detection of these P-like toxins provides support on the hypothesis of a possible “P-coding” system among the turrid snails. Similar approach can be done in other turrid species and the information utilized to correlate with their polychate-prey profiles.

#### ACKNOWLEDGMENT

ALH would like to thank Ms. Judy Ann Cocadiz for her invaluable support and contribution in preparing the final manuscript.

#### REFERENCES

Becker S. & Terlau H. (2008). Toxins from cone snails: properties, applications and biotechnological production. *Appl Microbiol Biotechnol*, 79:1-9. DOI: 10.1007/s00253-008-1385-6.

Fu Y., Li C., Dong S., Wu Y., Zhangsun D., & Luo S. (2018). Discovery methodology of novel conotoxins from *Conus* species. *Mar Drugs*, 16(11): 417. DOI: 10.3390/md16110417.

Gonzales D. & Saloma C. (2014). A bioinformatics survey for conotoxin-like sequences in three turrid snail venom duct transcriptomes. *Toxicon*, 1-9. DOI: 10.1016/j.toxicon.2014.10.003;

Heller J. (2015) Predators. In: *Sea Snails*. Springer, Cham. DOI: 10.1007/978-3-319-15452-7\_8;

Heralde FM (2007). Turrid molecular phylogeny, toxinology and feeding ecology. PhD Dissertation. UP-Diliman;

Heralde FM (2008). A rapidly diverging superfamily of peptide toxins in venomous *Gemmula* species. *Toxicon*, 51:890-897;

Imperial J., Cabang A., Song J., Raghuraman S., Gajewak J., Watkins M., Showers-Corneli P., Fedosov A., Concepcion G., Terlau H., Teichert R. & Olivera B. (2014). A family of excitatory peptide toxins from venomous crassispirine snails: Using constellation pharmacology to assess bioactivity. *Toxicon*, 89: 45-54. DOI: 10.1016/

Jin A., Muttenthaler M., Dutertre S., Himaya S., Kaas Q., Craik D., Lewis R., & Alewood, P. (2019) Conotoxins: chemistry and biology. *Chem Rev*, 119: 11510–11549. DOI: 10.1021/acs.chemrev.9b00207

Kantor Y., Stahlschmidt P., Aznar-Cormano L., Boichet P. & Puillandre N. (2017). Too familiar to be questioned? Revisiting the *Crassispira cerithina* species complex (Gastropoda: Conoidea: Pseudomelatomidae). *Journal of Molluscan Studies*, 83: 43–55. DOI: 10.1093/mollus/eyw036;

Kendel Y., Melaun C., Kurz A., Nicke A., Peigneur S., Tytgat J., Wunder C., Mebs D., & Kaufenstein S. (2013). Venomous secretions from marine snail of the terebridae family target acetylcholine receptors. *Toxins*, 5: 1043-1050. DOI: 10.3390/toxins5051043.

Miles L., Dy C., Nielsen J., Barnham K., Hinds M., Olivera B., Bulaj G., & Norton R. (2020). Structure of a novel P-superfamily spasmodic conotoxin reveals an inhibitory cystine knot motif. *J Biol Chem*, 277(45): 43033-43040.

- Omaga C., Carpio L., Imperial J., Daly N., Gajewak J., Flores M., Espino S., Christensen S., Filchakova O., Lopez-Vera E., Raghuraman S., Olivera B. & Concepcion G. (2017). Structure and Biological Activity of a turreptide from *Unedogemmula bisaya* Venom. *Biochemistry*, 56(45): 6051-6060. DOI: 10.1021/acs.biochem.7b00485;
- Olivera B., Corneli P. S., Watkins M. & Fedosov A. (2014). Biodiversity of cone snails and other venomous marine gastropods: Evolutionary success through neuropharmacology. *Annu. Rev. Anim. Biosci.* 2:487-513.
- Robinson R. & Norton R. (2014). Conotoxin gene superfamilies. *Mar Drugs*. 12:6058-6101. DOI: 10.3390/md12126058.
- Watkins M., Hillyard D., & Olivera B. (2006). Gene expressed in turrid venom duct: divergence and similarity to conotoxins. *J Mol Evol.* 62:247-256.



# Mass Spectrometry and Proteomics as Emerging Technologies for Breast Cancer

Maritess D. Cation,<sup>1,2,3</sup> and Maria Cristina Ramos<sup>1,4,5</sup>

<sup>1</sup>The Graduate School, University of Santo Tomas, Manila 1015 Philippines;

<sup>2</sup>Institute of Chemistry, Academia Sinica, Taipei 11529 Taiwan;

<sup>3</sup>Faculty of Pharmacy, University of Santo Tomas, Manila 1015 Philippines;

<sup>4</sup>Research Center for the Natural Sciences, University of Santo Tomas, Manila 1015 Philippines;

<sup>5</sup>College of Science, University of Santo Tomas, Manila 1015 Philippines

## Email address:

mdcation@ust.edu.ph

## To cite:

Cation, MD; Ramos, MC.2020. Mass Spectrometry and Proteomics as Emerging Technologies for Breast Cancer. PJBMB. Vol. 1, No. 1, 2020, pp. 12-28. doi: 10.5555/pjbmb.ph.2020.01.01.12

Received: 07 16, 2020; Accepted: 10 27, 2020; Published: 12 02, 2020

**Abstract:** Breast cancer among women has shown a steady increase in incidence and mortality rates in the Philippines, and around the globe. To date, there are a few and limited biomarkers approved for diagnosis and target for therapy. Some tumor tissues do not express any valid biomarkers in clinical tests, and patients from this group are unlikely to respond well to hormone therapy. Here, we presented a comprehensive literature resources citing potential biomarkers found from omics-based assays. More importantly, we also presented a rich list of significantly expressed novel protein biomarkers found through mass spectrometry and proteomic analysis. By applying mass spectrometry technology, we can achieve deep and large proteomic profiles from cells and tissues. The latest developments in mass spectrometry and its application will bring a big impact in breast cancer research and drug discovery as we find novel proteins and its association to various pathways linked to the hallmarks of breast cancer.

**Keywords:** Proteomics, mass spectrometry, breast cancer, biomarker, tumor marker, proteogenomics, omics

## 1. INTRODUCTION

The World Health Organization during the World Cancer Day 2020 recently warns the public of probable spike in breast cancer estimated to affect more than 15 million women globally by 2030 [1]. In the Philippines and most developing countries in Asia, a slow but steady increase in the rate of incidence, mortality and breast cancer recurrence is manifested. Health experts around the world have successfully implemented different strategies to mitigate and control breast cancer. They encourage individuals to conduct self-examination of the breast, and to undergo an annual breast cancer screening test to assess risk, and screen for breast cancer before its onset [2]. Breast cancer today is treated according to the presentation of specific breast cancer biomarkers in each patient. Breast cancer is detected with very few validated biomarkers - estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2).

However, there are a few patients who do not show evidence of any biomarkers with these in clinical laboratory tests, and therefore do not have good prognosis. Another drawback affecting accuracy in current clinical laboratory diagnosis in cancer is the availability of limited tissue biopsy for examination. The tissue obtained depending on its size may or may not be sufficient for the numerous pathological tests to perform. The biopsy tissue samples would naturally have a heterogeneous form which may lead to poor conclusive information on the properties of the cancer as a whole. These challenges are now being addressed with modern technology and sophisticated instrumentation such as the mass spectrometry paired with software for computational tools. It is regarded as fast and emerging technology useful in discovering new and potential protein biomarkers for breast cancer.

### A. Standard Breast Cancer Screening Methods

The female breast is made of milk-producing glandular tissues, lobules, ducts, and fatty tissues [3]. A thorough self-examination can observe changes on the skin, around the breast area, and lymph nodes near the armpit. Visualizing the lump can be done by breast mammography using low-dose X-rays to identify irregularities in the breast density and mass formation in the breast. Breast ultrasound is a better alternative to mammography since the later although noninvasive is painful and is not well tolerated by other patients. Magnetic resonance imaging (**MRI**) is another sophisticated but not too commonly used pictorial imaging of the breast. MRI provides a more accurate breast cancer staging results than the first two methods [4]. The images derived from these imaging methods are inspected by technicians and medical doctors to describe the nature of the size of the lump, malignancy of the disease, or make recommendation for additional laboratory examinations. The way this examination is done even with experienced personnel may pose inaccurate diagnosis, carry the risk of false-positive results, and discrepancies due to the limitations of the imaging method itself, and the subjective judgment of the observer. Nevertheless, these imaging technologies combined with regular self-examination have contributed to reduce mortality among breast cancer patients.

If the preliminary breast cancer screening results show suspicious lumps, other test may be recommended and a biopsy will be done [5]. Biopsy is an invasive procedure which collects breast tissue samples by open surgery or by needle biopsy collection. Immunohistochemistry (**IHC**) is the standard assay platform for the pathology of tissue biopsy and surgical resection specimen. In breast cancer, IHC for ER, PR, and HER2 are performed using laboratory produced antibodies. IHC may arrive at varying results due to the heterogeneity of the breast tissue, choice of antibody, age of fixative and reagents, different manufacturers, and the kind of immunostaining methods applied. In situ hybridization (ISH) or fluorescence in situ hybridization (**FISH**) test is done to complement IHC test to measure HER2 more accurately when IHC results are negative or at the borderline in terms of HER2 scoring.

### B. Gene and Protein Biomarkers Associated to Breast and Cancer Development

The mammary gland undergoes three distinct stages in which hormones and growth factors play critical role in the complex biological process called mammatogenesis. At the instance of embryo development, the skin's ectoderm will form mammary placodes until puberty [6]. The first stage forms an immature mammary gland in a process called "ductal morphogenesis" which is regulated by ER and insulin-like growth factor 1(**IGF-1**) [7], [8]. Upon reaching puberty, the growth factor-like amphiregulin (**AREG**) becomes more abundant for proper development of the mammary epithelial cells. Ovaries secrete estrogen hormone to promote fat cells

to proliferate and deposit around the breast connective tissues thereby enlarging the size of the breast [9]. As a person matures physically, progesterone, estrogen and other growth hormones are involved in the development of the breast, and production of milk during pregnancy. As a female adult becomes pregnant, prolactin and progesterone promote milk production. The alveolar at this time switches to secrete more milk during the lactation period. As the weaning period ceases to stop, milk production and supply begins to decrease with reduced demand. At this point involution takes place to return the mammary glands back to its normal pre-pregnancy condition [10].

As the breast ages, there is a decrease in the ducts embedded in strands of collagen, depletion of mammary stem cells, and replacement of interlobular connective tissues and glandular epithelium tissues with fat. Early onset of lobular involution in premenopausal women have significantly decreased the risk of developing breast cancer compared to those with delayed involution [11]. These stages in the normal mammary gland development, functioning and aging all require numerous chains of biochemical events involved in several signaling pathways. When any of these biomolecules undergo uncontrolled changes or mutation it becomes contributory to the risk of developing breast cancer.

Breast cancer are diagnosed in clinics for detection and analysis of the status of abundance of hormone receptors. Breast cancer tumors found to have an over expression or high abundance of these protein biomarkers will be called **ER+**, **PR+**, or **HER2+**. If none of these three receptors are found in the breast cancer tumor then such disease is considered as triple negative breast cancer (**TNBC**). With such presentation or absence of these receptor, breast cancer is subtype D accordingly in clinics. Table 1 summarizes the intrinsic breast cancer subtypes into 4 major groups [12].

Table 1. Intrinsic breast cancer Subtypes of breast cancer according to receptor and gene detection

Group	Breast Cancer Subtype	Biomarker/s overexpressed
1	Luminal A	ER+, PR+/PR-
2	Luminal B	ER+, PR+/PR-, and HER2+
3	HER2 positive, nonluminal	HER2+
4	Triple-negative or basal-like	No known biomarker

Estrogen has long been implicated to support growth of breast cancer because it binds directly to ERs found on the cell membrane. Estrogen receptors are transcription factor proteins encoded by the estrogen receptor 1 (**ESR1**) gene. It is composed of domains which support hormone binding, and activation of transcription for the

synthesis of proteins. The ER+ breast cancer manifests abundant ERs that bind with estrogen promoting deregulated breast cancer growth. Anti-estrogen drug is an anti-estrogen hormone therapy acting as estrogen antagonist to block estrogen from binding to the ERs. Progesterone is an important steroid hormone which regulates changes in the reproductive system. Most breast cancers express both ER and PR together, as a result PR is usually studied always alongside with ER. The progesterone receptor positive (PR+) kind of breast cancer would not only be PR+ but often be ER+, too. Hormone receptor negative or TNBC do not over express ER, PR nor HER2 receptors. TNBC to date remains to have the worst prognosis and available hormone antagonistic therapies in most situations will not work with this type of cancer.

In a study by Chiang et al., they found that in ER+ invasive breast carcinomas, AREG was contributory to the increased invasiveness of breast cancer cells. Suppressing AREG expression in transformed human breast epithelial cells in nude mice showed decrease formation of tumor [13]. AREG mechanism have not been completely understood but it is hypothesized to cross the membrane to the stromal fibroblasts activating the epidermal growth factor receptor (**EGFR**), and induce expression of fibroblast growth factors (**FGFs**) leading to cellular proliferation. In adult female breast, some hyperplastic enlarged lobular unit (**HELU**) studies revealed both ER and AREG are upregulated and are suspected to induce self-propagating growth leading to tumorigenesis [14]. Such findings support that AREG plays a role in breast cancer development particularly for ER+ tumors [15], [16].

Another study identified progesterone and its relation to the tumor necrosis factor ligand superfamily, member 11 gene which encodes the receptor activator of nuclear kappa-B ligand (RANKL) and its receptor – RANK. RANK, RANKL and progesterone are all found to be essential in the milk production of the mammary gland during pre-pregnancy stages [17]. The RANK/RANKL signaling pathway is also regulated by progesterone being the main upstream regulator of the RANK/RANKL pair. The RANK/RANKL pair expression also were found to control stem cell expansion and generate the pro-growth response, and drives cell proliferation in progestin-dependent breast cancers [18], [19].

Risk of breast cancer was investigated from isolated extracellular matrix (**ECM**) from post-weaning mammary glands. The study revealed some mass forming ECM fragments which increases the metastasis of breast cancer in mice, and the invasiveness of breast cancer cells in culture [20]. Post-lactation involution processes have also been connected to breast cancer and tumorigenesis in few transcriptional profiling studies [21]. It is assumed that involution may lead to a tumor microenvironment that alters preneoplastic mammary cells leading to transient increase in breast cancer risk after pregnancy [22].

A few of the linked oncogenic pathways to breast cancer are the ER signaling pathway [23], HER2 signaling pathway [24], nuclear factor-kappaB (NFkB) signaling pathway [25], mitogen-activated protein kinase (MAPK) signaling pathway [26], phosphatidylinositol 3-kinase/protein kinase B/mammalian target of rapamycin (PI3/AKT/mTOR) signaling pathway [27], and notch signaling pathway [28]. When sudden changes in the body's physiology and conditions take place, these signaling pathways maybe altered, hijacked or dysregulated by activities of the cells leading to uncontrolled cell growth, cell invasiveness, and suppresses cell apoptosis [12]. These pathways are considered in most breast cancer studies and some drugs have been developed to inhibit downstream signaling to control breast cancer.

Other studies look at the genetic DNA for preventive and prophylactic approach to breast cancer treatment. Germline or somatic mutations in genes are suspected to code for growth of the cancerous cells [29]. One of the genes linked to breast cancer is the human epidermal growth factor 2 gene, also called HER-2/neu or ErbB-2 gene (**ERBB2**). The amplification and deregulation of the ERBB2 gene leads to an overexpression of HER2 in tumor cells. Burstein et al., found out that HER2 overexpression may reach as high as 50-fold higher than its normal abundance in 30% of the invasive breast cancer tumor tissues used in this study [30]. The HER-2 protein is found to activate molecular pathways supporting cellular proliferation and metastasis of breast cancer.

Some breast cancers are hereditary and it is passed from an autosomal dominant pattern from any one of the parents to offspring such as the mutations presented by the breast cancer gene 1 (**BRCA1**) and breast cancer gene 2 (**BRCA2**) [31]. Presence of these genes increases one's risk to developing breast cancer and morbidity [32]. At normal conditions, the breast cancer genes function by encoding proteins to control and repair damages in the DNA or fix other gene mutations. BRCA genes also prevent uncontrolled cellular division and growth suppressing rapid tumor formation. But when BRCA genes are mutated it leads to forming abnormal and nonfunctional versions of the corresponding BRCA1 and BRCA2 proteins.

From the abovementioned biomarkers still very few are currently being considered for breast cancer clinical diagnosis to date. These biomarkers are either genes or proteins, with known cellular regulation and binding properties controlling downstream signaling activities. In some instances, these molecules are modified by external factors leading to its distorted structure affecting its role to maintain the cell's normal homeostasis. With increasing demand for potential biomarkers for diagnosis, a high throughput and multiplex analysis method in the detection is an emerging need. Biomarker discovery by **mass spectrometry (MS)** is becoming a popular method of

choice, and as well as an effective tool in disease diagnosis. This method is capable of analyzing tissue samples that are heterogeneous in nature, and are usually available in small or limited quantities. Experienced MS users were able to achieve wider and deeper analytical capability in many different kinds of samples. MS has also delivered analysis with high sensitivity, reproducibility and repeatability, and accuracy and precision over multiplex analysis of samples. In MS, profiling of the biopsy of both tumor (T) and its adjacent non-tumor tissues (NT) can also be done simultaneously. This method can deliver better identification and differentiation measures of breast cancer-associated biomarkers in every patient.

### C. Genes to protein biomarkers in breast cancer clinical diagnosis

Prognosis estimates the course and outcome of cancer such as the likelihood of recurrence, remission and survival. An individual breast cancer prognosis is done in several phases, first tumor biopsy is subjected to histomorphology to examine its type, grade and size of the cancer, and its presence in the lymph nodes. The second phase detects protein biomarkers, ER, PR, and HER2 expression status. In cancer prognosis, about 20-30% of breast cancer tumors have malignant breast cancer cells that have unusually high concentration of HER2 receptor proteins. Overexpression of HER2 found on tissues are contributory to unusually rapid cell proliferation. This is carried on by the dimerization of HER2 with other EGFRs leading to the activation of the growth factor signaling pathway driving cancer to grow [33]. Some HER2 breast cancer patients show improvement with first generation adjuvant therapy like trastuzumab, while others would require other HER2 targeted therapies such as antibodies pertuzumab and adotrastuzumab emtasine, or a kinase inhibitor like lapatinib [34].

Another protein biomarker used in breast cancer care is **Ki-67** antigen expressed by the marker of proliferation Ki67 (**MKi67**) gene. This is used as a prognostic biomarker for measuring proliferation, predicting drug response, and drug resistance [35]. Ki67 concentration in tumor positive cells increases as cells prepare to divide and form new cells. Hormone receptors together with Ki-67 are quantitatively measured using IHC based assays [36]. Results from the IHC assay have brought some concerns such as the nonlinear nature of IHC staining due to the heterogeneous nature of the tumor tissue, the antibodies itself and slide scoring applied, and the different subcellular location of the different biomarkers. Genomics have addressed these issues and led to the development of various gene-based technologies for tumor biomarker assessment. There is a free online tool called "Predict," that may be used to provide patients and doctors varied treatments options for post-operative early invasive breast cancer patients [37]. Estimating prognosis have also been successfully done with

genomics assays approved by USFDA to profile gene expression and has been used as one of the more accurate assays for diagnosis and estimating prognosis in breast cancer. In various conditions and stages of breast cancer, several multianalyte tests are used routinely in the laboratory examination of patient samples. Current gene-based assays include *Oncotype DX*, *urokinase plasminogen activator (uPA)-PAI-1*, *BBDRisk Dx*, *Immunohistochemistry 4 (IHC4)*, *BreastSentry*, *MammaPrint*, *EndoPredict*, *Breast Cancer Index*, *Natera's Signatera Molecular Monitoring (MRD)*, *Rotterdam Signature 76-Genes Panel*, *HERmark*, *NexCourse IHC4*, *Mammostrat*, *Symphony (Agendia)*, *GeneSearch BLN*, *Insight TNBC type*, and *Prosigna (PAM50)*. Each assay has its own advantages and applies only to specific subgroups of patients to predict prognosis and plan for the most suited adjunct therapy treatment.

There are also individual genes USFDA approved for prognosis in some types of breast cancers such as *neurotrophic tropomyosin receptor kinase (NTRK)* genes, and *phosphatidylinositol 3-kinases catalytic subunit alpha (PI3KCA)* gene. NTRK genes were observed upon the fusion of proteins encoded upon by three NTRK genes in secretory breast carcinoma [38]. The PI3KCA gene commonly found mutated in ER+ breast cancer alters the activity of class IA phosphatidylinositol 3-kinase (**PI3K**). The mutation causes PI3K to downstream activate the PI3K/AKT/mTOR pathway involve in breast cancer [39]. Patients with advanced breast cancer having PI3K mutation are given some PI3K inhibitor like alpelisib along with fulvestrant [40]. These findings suggest that detecting genes is useful in determining therapies for lowering the risk of breast cancer onset and recurrence. On the other hand, incorporating genetic testing into breast cancer care requires accredited laboratories to perform this clinically validated tests, technical skills of analyst, and modern instrumentation to arrive at accurate results.

Complementing genomics is the emerging proteomics approach to diagnosis and prognosis. The study of proteomics has brought remarkable advancement in the better understanding of the disease since proteins are directly involved in all cellular processes. The carcinoembryonic antigen (CEA) family of glycoproteins is the most widely used antigen found in the tumor tissue. One of the major subfamilies of these genes is the **CEA** cellular adhesion molecule (**CEACAM**) family belonging to the immunoglobulin superfamily. The adhesion properties of these molecules with one another or with other molecules suggest that alternations in cell adhesion play an important role in cancer metastasis. CEA levels are measure to aid in breast cancer diagnosis, clinical staging, responsiveness to chemotherapy or radiotherapy treatments, and monitoring recurrence in post-operative patients [41]. The cancer antigen 15-3 (**CA 15-3**), and cancer antigen 27.29 (**CA 27.29**) are normally

expressed in healthy cells to work in controlling abnormal cellular growth. In cancer cells, elevated CA 15-3 amounts in blood are used to monitor the stage of breast cancer, and measure the effectivity of a breast cancer therapy [42]. A more sensitive biomarker than CA 15-3 for metastatic breast cancer is CA 27.29. This highly polymorphic glycoprotein is found to be less glycosylated in breast cancer tumor cells than its usual form [43]. The biomarker is expressed throughout malignant epithelial cells of the breast and measured together with other tumor markers to improve specificity in disease staging, track effectivity of therapy, and monitor breast cancer recurrence.

USFDA have approved protein tumor markers for breast cancer longitudinal study including the circulating tumor cell analysis of epithelial cell adhesion molecule (**EpCAM**), **CD45** antigen, and cytokeratins (**CK8**, **CK18**, and **CK19**). Circulating tumor cells (**CTCs**) though exist in minute concentrations in the blood, urine, stool or other body fluids have occupied a big role in many cancer research seeking to find prognostic and therapeutic values in metastatic breast cancer [18], [44]. EpCAM expressed by normal CTCs, is a membrane protein being considered to become a novel drug target for gene therapy. In a study done by Osta et. al, they compared the primary and metastatic breast cancer against normal breast tissue and discovered that EpCAM mRNA expression levels to be differentially overexpressed by 100 to 1000-fold in breast cancer tumor tissue [45]. Such tumor markers may also be measured periodically to monitor good response of cancer therapy by detecting a significant decrease in the concentration of circulating tumor marker. **CD45** antigen also known as protein tyrosine phosphatase, receptor type, C (**PTPRC**) is a signaling protein involved in a number of cellular processes including cellular division and growth. It is also found to regulate functions of some antigen receptor complexes and kinases necessary for antigen receptor signaling [46]. In breast cancer CD45 are important in the diagnosis and can be used to monitor changes relative to the effectivity of the applied drug for therapy. CK8, CK18, and CK19 are proteins expressed by the epithelial cells normally lining the breast tissue. Measuring changes in the levels of these protein biomarkers concentration would determine the cancerous condition of breast-associated adenocarcinomas [47], [48]. Lustberg *et al.* also reported in their findings that circulating atypical cells detected in blood have high expression of CK8, CK18, CK19, and CD45 biomarkers [49]. Researchers continue to discover possible mechanisms of these CTC biomarkers relevant in understanding tumor metastasis. The first and only approved immunotherapy currently used in breast cancer clinics is the detection of **programmed cell death-1 receptor (PD-1)**. **PD-1** is an important signaling protein found on the surface of immune cells and some tumor cells Its role is to provide immune inhibitory signals to fight pathogens and cancer cells. By blocking PD-1 with drugs such as

atezolizumab, PD-1 can help boost the body's immune system [50]. In some forms of cancer, IHC analysis is performed to detect **PD-1 ligand (PD-L1)** overexpression from patient samples. The JAK1/JAK2-STAT1/STAT2/STAT3-IRF1 pathway regulates the expression of PD-L1 upon secretion of interferon gamma (IFN- $\gamma$ ) by the T cells [51]. The molecule PD-1 once bound to PD-L1 delivers a negative modulatory signaling pathway to activate T cells. The activation of due to PD-1 and its ligand PD-L1 in the PD-1/PD-L1 pathway induces downregulation of T-cell activity, cell proliferation, induction of tolerance to antigens, and trick the immune cells from destroying damaged cells [52]. In metastatic TNBC, blocking PD-1 or PD-L1 with specific antibodies have shown good response and control over cancer growth [53]. IHC when compared to MS-based analysis of PDL1, IHC was found to show lower concentrations of the glycosylated protein. This result may lead clinicians to ignore PDL1 expression and may misdiagnosis patients with inappropriate immunotherapy. In this respect, MS-based absolute quantification delivers better sensitivity to detect protein expression levels even when it is in its glycosylated form [54].

Proteomics and genomics, together termed as proteogenomics, in conjunction with computational tools, newer technologies are leading the way to finding better solutions in discovering specific biomarkers to combat breast cancer. This underscores as well the importance of biomarker in the more advanced stages, recurring types and those that are rare, and have difficulty in finding appropriate and precise treatment. Collectively, these attempts in proteogenomics is leaning towards an improved quality-of-life among breast cancer patients, and an eradication of the disease in the near future.

#### **D. Mass Spectrometry-based Proteomics Approach**

Research in breast cancer has not found a biomarker which provides measure for accurate breast cancer diagnosis and typing. Protein molecules are considered best biomarkers for they have different functions and are involved in many biochemical reactions in the body including the body's fight to prevent diseases including cancer. To perform analysis of the proteins, mass spectrometry was widely adopted for proteomics research. In the past years, peptides cannot even be analyzed in MS since peptide samples cannot be ionized in its gaseous form. With the latest application of ionization technology even **peptides** and other simple molecules to as low as femtomole quantities in liquid states can possibly be analyzed and sequenced today in MS.

MS-based proteomics strategies can either be top-down, or bottom-up proteomics [55]. Proteins are first extracted from biological samples like cell lines, tissues, tumor specimens, and other derivatives. These proteins are enzymatically digested into peptides and then analyzed by liquid chromatography-tandem mass

spectrometry (**LC-MS/MS**). In **top-down proteomics** it analyzes intact proteins separated first by isoelectric focusing (IEF) according to its isoelectric pH, then by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) according to its size. The separation of proteins can be viewed by performing protein staining or by using fluorescent tags. The two-dimensional electrophoresis (2-DE) technique is low throughput and is limited by protein solubility and sensitivity of detection. In this technique a big amount of sample is needed for this assay. Its minimum detection sensitivity is down to a few numbers of proteins with molecular weights of at least 120 kDA per run. Specific proteins of interest are excised from the gel, purified, and analyzed using the mass spectrometer [56]. An improved protein profiling strategy was later developed called the “**bottom-up**” or “**shotgun**” **proteomics** [55] which refers to reconstruction of protein information from peptide sequences. In shotgun proteomics high performance liquid chromatography (**HPLC**) in tandem mass spectrometry is used for characterization of proteins from complex samples [57], [58], clinical biomarkers discovery [59], identification of post-translational modifications (PTMs) [60] and protein-protein interactions [61] to explore biological system [62].

Peptides are digested proteins extracted from samples and digested usually with trypsin, or other proteases. A final step requires peptide purification by washing away salts, detergents and inactivated enzymes to eliminate matrix effects in LC-MS leading to ion suppression or co-elution of distinct peptides. Purified peptides are analyzed with optimized workflows and LC-MS method settings to achieve efficient mass fragmentation datasets for identifying peptide sequences and proteins. The resulting analysis will be matched to protein sequences from protein sequence databases [63] and spectral libraries of fragment ions and precursor ions [64], [65]. Spectral libraries are acquired a priori from single shot and fractionated samples from tissues or cell lines from an organism. The spectral libraries are used for peptide identification and quantification for protein sequencing at 1% false discovery rate [66]. MS engineering technology have continuously achieved better results over the years. Latest chromatography columns have been designed with varying dimensions, retention characteristics, particle size, and selectivity of the column material. These enhancements make it possible to detect a broader protein range, better peak resolution and shorter gradient time. Using LC-MS technology proteome profiling acquired from reproducible analysis have presented improved sensitivity in peptide and protein identification and quantitation, specificity and discrimination between protein isoforms. Datasets derived from MS analysis of the peptides are submitted to software capable of automatically sequencing the MS/MS (also written as MS<sup>2</sup>) spectra of the fragmented ions. Peptide identification is done by comparing the mass spectrum

of peptides to mass spectra predicted from public sequence databases or mass spectral libraries.

The development of quantitative methods has become the focus of MS-based proteomics research, with the aim of achieving high precision and accuracy in quantitation, as well as a high reproducibility and low number of missing values. Protein quantification is usually performed in two major approaches: the use of stable isotope labelling and label-free techniques. These methods provide opportunity for sample multiplexing, and quantitation analysis based on relative intensities of the reporter ions. Most proteomics quantification workflows involve chemical labelling techniques for relative and absolute quantitation, such as Stable Isotope Labeling by Amino Acids in Cell Culture (**SILAC**), Isobaric Tags for Relative and Absolute Quantification (**iTRAQ**), and Tandem Mass Tags (**TMT**). Labelling strategies allow multiplexing few samples at the same time, however, they are limited by the high cost of isotope labels, varying labeling efficiency and software for data analysis. Some of the techniques of labeling include protein or peptide, chimeric recombinant protein, isobaric and metabolic either in vivo or in vitro. SILAC [67] is used in vivo, while iTRAQ [68], and TMT [69] are used in vitro.

SILAC requires cells cultured with light medium using normal arginine (blue color Arg-0 isotope), and medium with heavy arginine (red color Arg-6 isotope). These isotopes of arginine are metabolically incorporated into the proteins while cells are growing under certain experimental conditions. Afterwards, cells are harvested and proteins are extracted from each set-up. Mass spectra of the corresponding peptides in both medium and light cultures are analyzed. Combination of both light and medium sample peak intensities in the mass spectrum will give the ratio of its relative protein abundance. If the protein would have a ratio of 1 this means that the abundance in both light and medium samples are the same. If the protein ratio is less than 1, then the protein abundance is greater in the medium sample than in the light sample, and vice versa if ratio is more than 1. These information from SILAC may lead to the identification of differentially expressed proteins in the sample. iTRAQ is also used for cells, tissues and other samples. It uses stable isotopes in iTRAQ reagents to be covalently bonded to the N-terminus and side chain amines of proteotypic peptides. iTRAQ enables relative quantification of very complex tissue mixtures thus it is been used in MS analysis in various applications such as comparison between normal and adjacent tissues, drug treated samples of cancer patient samples, biological replicates. TMT is currently the latest add-on in the label-methods for MS/MS analysis and have great advantage for its ability for multiplex analysis surpassing that of what iTRAQ may offer. TMT can perform analysis with several samples within one experiment producing MS data sets that defines the proteome of each sample. By TMT method analysis can have a reduced overall analytical time and eliminates

variations from one sample run to another. For the protein quantitation and identification, the isobaric set of six mass tags with five isotopic substitutions called tandem mass tags six-plex (6-plex-TMT) were the first to be used as chemical labels for cell lines or patient samples. There are different TMT products now made available capable of conducting as high as 16-multiplex (16-plex Pro-TMT) analyses. TMT method was also compared to label-free DIA method in a study performed in the lab of Muntel [69]. They found that MS label-free approach yields high protein quantification and identification with less than 2% missing values, while TMT approach have achieved higher quantitation accuracy.

These labelling approach to proteomics improve sensitivity and the analysis when compared to other approach display fewer missing fragment ion intensities in the analysis. Table 2, highlights the characteristics of the three labeling techniques.

Table 2. Characteristics of SILAC, iTRAQ, and TMT

	SILAC	iTRAQ	TMT
Labelling Method	Stable isotope - labeled lysine and arginine	Isobaric reagents	Stable isotope labels
Sample	Living cells	Peptides	Proteins and peptides
Advantages	Less sample required Can be used in live cell lines	High throughput (4-/8-plex) High sensitivity Good reproducibility	High throughput (up to 10-plex) High sensitivity Efficient separation ability Good reproducibility and repeatability
Disadvantages	Expensive reagents Limited to cell line samples	Expensive reagents	Expensive reagents Low scan speed requirement decreases sample throughput
Reference	[67]	[70]	[71]

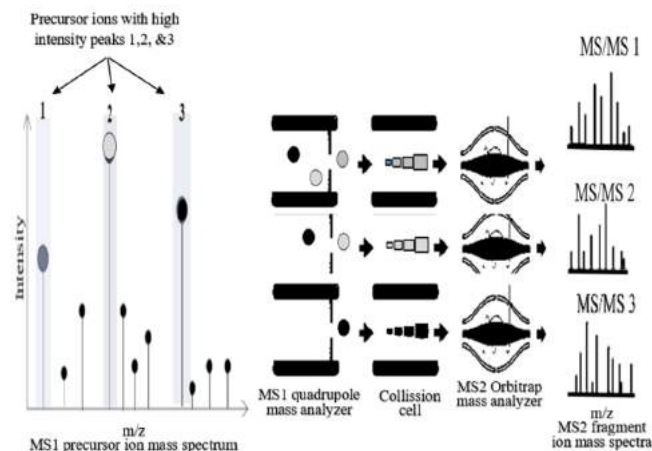
MS-based proteomics strategies make it also possible to perform measurement of proteins that have undergone some form of PTMs including ubiquitylation, phosphorylation, and acetylation to name a few among hundred types [72]. PTMs enable energy-efficient protein function optimization and diversifying its functionality at the cellular level to regulate cellular processes and pathway signals for growth, proliferation, and apoptosis. In a pool of peptides, PTMs are mapped by shotgun sequencing, or are fished out with a variety of affinity-enrichment or by covalent capture techniques, or its combination. In doing so, specificity is improved and co-eluting distinct peptides prevented. An approach to improve shotgun sequencing approach in LC/MS is by performing a prefractionation step. This will improve the dynamic range and expands the LC/MS-MS capacity to detect beyond the four to five orders of magnitudes [73]. By doing so, an increase in peptide coverage and identification of low abundant proteins (LAPs) to the high abundant proteins (HAPs) [74]

happen. Protein abundance in some samples like in a single cell protein, can vary from 50 to 1,000,000 relative quantities. When the HAPs and LAPs are together, proteomic analysis becomes difficult since the bigger molecules can dominate in the sample matrix and hide the appearance of LAPs. There is also a big consideration in research of LAPs for some studies suggesting that these are found to leak into the blood in very small quantities. Circulating tumor cells are known to contain LAPs which may yield potential breast cancer biomarkers.

PTM-specific enrichment approach improves sensitivity and specificity of MS analysis [75]. In this manner, phosphoproteins are isolated from proteomic mixtures either by anti-phosphoamino acid antibodies or PTM-specific affinity to the phosphate groups. After targeted enrichment, these phosphoproteins are digested to yield peptides which is later enriched by antibodies, titanium dioxide (TiO<sub>2</sub>), or by immobilized metal affinity chromatography (IMAC). At this point peptide complexity and heterogeneity is addressed by separating it into components in ion exchange chromatography as final step prior to LC-tandem MS analysis. The result of this leading to the detection of thousands of phosphorylated sites [76].

Samples may also be processed without using isotopic labels because of some practical reasons and cost-consideration. In label-free MS, peptides are directly analyzed, and proteins are quantified on the basis of precursor ion signal intensity or spectral counting [80]. Recently, label-free quantitative proteins strategy has become a stand-alone method or in combination with enrichment or other labelling methods. In label-free approach, no chemical labels or tags are added to peptides before it is submitted to MS analysis. Fragmentation and mass analysis of these peptides are obtained either by data dependent (DDA) or data independent acquisition (DIA) [77]. **Data dependent acquisition (DDA)** is a highly selective method commonly applied to many MS instruments including triple quadrupole, Orbitrap, and tandem Q-ToF or ToF/ToF. Large-scale precursor ions enter the quadrupoles. The quadrupole emits simultaneously high and low energies to facilitate collision-induced dissociation of precursor ions. Upon fragmentation, accurate mass measurement is detected in the instrument. In DDA, the peptide precursors are scanned and in each scan about 9-10 precursors having the highest intensities at MS1 are selected for further fragmentation for the sequential MS/MS. However, in DDA method only those with high intensity precursors are selected to enter the mass analyzer. The low intensities fragments are ejected out in a trajectory and will never reach the detector for measurement. DDA may eliminate some important peptides at this stage. It may also miss some precursors to be identified after MS/MS analysis creating some difficulties in quantitation.

The **data independent acquisition (DIA)** method is an answer to the earlier difficulties faced by using DDA. In DIA the entire mass spectrum produced from MS1 ion intensities is fragmented and undergoes MS/MS analysis. Precursors are taken part by part selected over a very narrow  $m/z$  range. This is known as isolation windows, and the acquisition method termed as sequential windowed acquisition of all theoretical

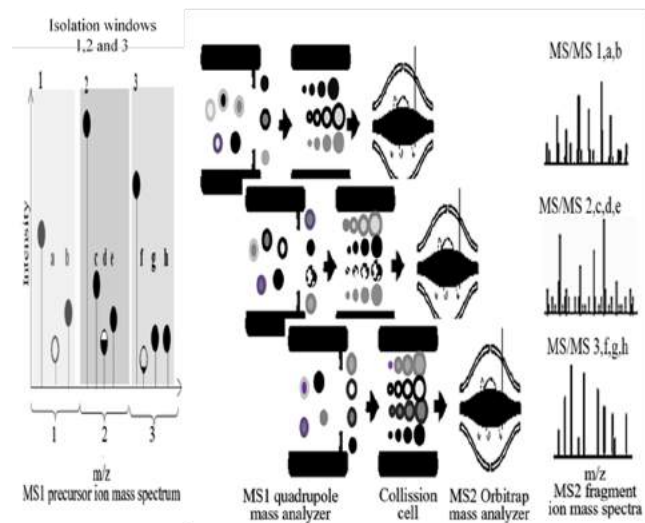


**Figure 1. Data Dependent Acquisition.** In the figure the high intensity precursors 1,2,3 in MS1 full scan will undergo sequential isolation and MS/MS fragmentation to produce a mass ion spectrum. (Abbreviation: HCD, High-energy collision dissociation)

fragmented ion mass spectrometry (**SWATH MS**) [77]. It is a high throughput label-free proteotyping technique applied in the mass analysis of every precursor ion found within narrow isolation windows. The precursor ions are fragmented over pre-selected smaller ranges of precursor  $m/z$  ratio range. All precursor ions enter sequentially the collision cells before reaching the final mass analyzer.

Due to the highly complex and convoluted DIA MS/MS spectra, it is difficult to interpret for peptide identification by conventional database searching tools without pre-processing. Instead the data analysis mainly relies on a prior knowledge information derived from fragment ion spectra of the targeted peptides. Commonly, there are two primary ways to interpret MS DIA data: (1) peptide-centric analysis (library query targeted analysis methods) and (2) spectrum-centric analysis (library-free approaches) [78]. Peptide-centric analysis needs a prior-knowledge information “peptide query parameters (PQPs)” stored in a spectral library. The library contains information of each targeted peptide consists of precursor  $m/z$ , peptide sequence of fragment ions, highly confident fragmented ions, and their relative intensity, and the normalized retention time. A high-quality and comprehensive **spectral library** is required for MS label-free identification and quantification. The

spectra, it is difficult to interpret for peptide identification by conventional database searching tools without pre-



**Figure 2. Data Independent Acquisition.** In the figure all precursors produced in MS1 full scan will undergo MS/MS fragmentation in a sequential narrow  $m/z$  isolation windows 1,2 and 3.

processing. Instead the data analysis mainly relies on a prior knowledge information derived from fragment ion spectra of the targeted peptides. Commonly, there are two primary ways to interpret MS DIA data: (1) peptide-centric analysis (library query targeted analysis methods) and (2) spectrum-centric analysis (library-free approaches) [78]. Peptide-centric analysis needs a prior-knowledge information “peptide query parameters (PQPs)” stored in a spectral library. The library contains information of each targeted peptide consists of precursor  $m/z$ , peptide sequence of fragment ions, highly confident fragmented ions, and their relative intensity, and the normalized retention time. A high-quality and comprehensive **spectral library** is required for MS label-free identification and quantification. The spectral library is built from actual MS experiments in DDA mode acquisition and processed in a software like Maxquant [79] and Spectronaut [80] to generate a database of mass spectra. To achieve a normalized retention time, a set of shared endogenous or synthetic spike-in peptides such as the indexed retention time (**iRT**) are spiked in sample for accurate prediction of peptide retention time. This creates MS datasets with consistent, comprehensive and validated digital map of the entire proteome. False discovery rates are set to control the propagation of errors in MS analysis [81]. Some organism scale spectral libraries or are now publicly available include *Homo sapiens*, *Methylobacterium extorquens* (strain PA1), *Drosophila melanogaster*, *Solanum lycopersicum* and *Streptococcus pyogenes* to mention a few [82]. To date, spectral libraries are continuously being expanded to include other organisms and specific human tissues to achieve deeper proteomic coverage.



As DIA is widely being used at present in MS data acquisition simultaneous development of software tools aided by artificial intelligence (AI) are also in progress. Common software like OpenSWATH [83], Skyline [78], and MSPLIT-DIA [84] have enabled processing and simplified the analysis of highly convoluted SWATH-MS data. On the other hand, PQPs information can also be computationally predicted in silico spectral libraries by using deep neural networks to predict MS/MS spectrum, retention time, and fragment ion intensity. These in silico spectral libraries are used to interpret DIA data in deep learning methods such as DeepMass [85], Prosit [86], and DeepDIA [87]. These peptide-centric analysis approaches are sensitive and have more comprehensive DIA data than what is normally found in sample-specific generated spectral libraries. However, they limit peptide identification only to analytes found present in silico libraries. Thus, tools designed to detect peptides from DIA data without libraries will soon be introduced.

In spectrum-centric analysis, DIA spectra are most commonly interpreted using classical database search strategies [88]–[90]. This approach detects precursor ion chromatographic features and deconvolves the DIA fragments into pseudo-MS/MS spectra, which can then be directly searched with the traditional sequence database. Current spectrum-centric analysis DIA data software tools include Pulsar in Spectronaut [80], Group-DIA [91], PECAN [92], PEAKS [87], and DIA-Umpire [93]. However, these library-free based tools are not as sensitive as the library-based approaches [94], although these strategies have its advantages in identifying new peptide variants in DIA data sets. Further, spectrum-centric library-free search has demonstrated its potential in identifying novel peptides. To date, most published studies are still using the spectral library-based targeted extraction approach [94].

### E. Proteomics for breast cancer molecular subtyping

Biomarkers have tremendously increased over the past decades, due to the advancement of high-throughput platforms for investigating the molecular characteristics of tumor tissue biopsies. With powerful bioinformatics tools integrated to aid in breast cancer research, it enabled us to further dissect the tumor to its molecular level and expand the spectrum of breast cancer subtypes. Microarray-based gene expression profiling has helped in determining breast cancer from its histopathologic type to its molecular subtype [95]. Today, ER+ and ER- breast cancer subtypes are considered as different diseases and are treated depending on the presentation of significant biomarkers [96]. Due to these development in genomics research on breast cancer biomarkers, The Cancer Genome Atlas (TCGA) Network has refined subtypes of breast cancer based from the extensive profiling of protein expression levels, microRNAs, and DNA gene

mutations [97]. The molecular subtypes luminal A, luminal B, HER2-enriched, and basal-like breast cancers (Table 1) are undergoing paradigm changes at the molecular level to help in the modernization of breast cancer treatment [97]. Early proteomic studies of clinical breast tumor samples got low proteome coverage yield from low cohort sizes. Rezaul et. al, have identified about 1000 proteins from each of the 6 patient samples from which they found more than 200 differentially expressed proteins between the **ER+** and **ER-** breast cancer tissues. Their study discover the potential biomarkers **Fascin, death-associated protein 5, Iprin- $\alpha$ 1, and  $\beta$ -arrestin 1** specifically found expressed only in the **ER-negative subgroup** [98]. Cha et al., studied 18 breast cancer samples of different subtypes and detected 298 significantly changing proteins that are associated to the transitional changes from the normal epithelial tissue conditions to a highly invasive malignant tumor. In their work, they found several proteins which are involved in alterations of downstream transcription factors (TFs) affecting regulatory pathways in breast cancer which can be used distinguishes **malignant tumor from matched normal tissues** [99].

With ongoing improvement in MS technologies, sample preparation, and bioinformatics tools have extended MS capacity to analyse larger cohort sizes. These developments led to improve both quality and quantity of proteomic data. In the lab of Liu et al., they made use of laser capture microdissection–nanoscale LC–MS/MS approach to study fresh frozen paraffin embedded breast tissues from 126 **TNBC** breast cancer samples. They were able to quantify about 3500 proteins, and identified 10 upregulated- potential protein biomarkers for novel therapies of distant metastatic TNBC. The study found that these proteins are involved in cell metabolism, cell death, immune response, transport of macromolecules, and biological processes linked to cancer progression. These proteins are apoptosis-inducing factor 1, mitochondrial (**AIFM1**), AP-1 complex subunit gamma-1 (**AP1G1**), AP-1 complex subunit mu-1 (**AP1M1**), F-actin-capping protein subunit beta (**CAPZB**), UMP-CMP kinase (**CMPK1**), catenin alpha-1 (**CTNNA1**), echinoderm microtubule-associated protein-like 4 (**EML4**), ferritin heavy chain 1 (**FTH1**), **GANAB**, , and syntaxin-12 (**STX12**) [100].

De Marchi et al., profiled **ER+** breast cancer patients, by comparing primary tumors with and without lymph node involvement to their matched normal noncancerous tissues. Their study contributed to the understanding of important functional insights of proteins expressed among the ER+ tamoxifen resistant population. The analysis has obtained more than 9000 proteins, where they were able to identify four important proteins namely, cingulin (**CGN**), Ras GTPase-activating protein-binding protein 2 (**G3BP2**), the programmed cell death protein 4 (**PDCD4**), and ovarian carcinoma immunoreactive antigen domain-containing protein 1 (**OClAD1**). They suggested these to be potential

biomarkers in predicting prognosis for the tamoxifen-susceptibility in recurrent breast cancer [101].

Tyanova et al., used SILAC MS labelling approach in the analysis of luminal **ER+**, **HER2+**, and **TNBC** tumor subtypes. Their team used support vector machine (SVM)-based classification for the 10,000 protein groups they identified from 40 tumor tissue samples. The functional proteomic profiles from these samples were able to distinguish one subtype from another such as proteins related to cell growth, cell-cell communication energy metabolism, and mRNA translation. From this analysis, they identified **19 specific proteins** to differentiate between the breast cancer subtypes [102].

Some studies have found integration of genomics and proteins effective in breast cancer subtyping. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) breast cancer study with 77 breast cancer samples from the TCGA cohort [103] accurately quantified a total depth of about 11,000 proteins from combined method of iTRAQ labelling approach and fractionation of samples. Comparison between the **intrinsic breast cancer subtypes** with the mRNA intrinsic subtypes found relatively similar abundances in the hormone receptors ER, PR, and HER2, phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform (**PK3CA**), GATA-binding 3 (**GATA3**), and cellular tumor antigen p53 (**p53**). They were also able to show an unsupervised clustering of samples based on the proteomic level and identified these into three main clusters – luminal enriched, basal enriched, and stromal enriched. Their findings revealed that very similar subtype-defining features can be observed in both RNA-seq and labeled MS protein analysis although different tissue sections of the same tumors were used [103].

Johansson et al., by applying nanoLC-MS/MS method have accurately identified about 14,000 proteins, and 13,000 genes [104]. From these data 9995 of these genes identified across all 45 breast cancer samples were used for unsupervised clustering and arrived at 6 distinct proteome-based consensus core tumor clusters (**CoTC**). These clusters were closely associated to previously determined subtype in 50 transcripts (PAM50). Using the PAM50 subtyping, the clusters identified are **CoTC1** and **CoTC2** (basal-like tumors), **CoTC3** (luminal A), **CoTC4** and **CoTC6** (dominated by luminal and HER2+ tumors), and **CoTC5** (normal-like breast cancer). On the other hand, the depth and quality of proteome profiling of these clusters led to identify novel immunohistochemical biomarker candidates can help achieve better patient stratification approach to treatment. Moreover, their discovery also identified the link between tumor extracellular matrix composition to immune cell infiltration to prognosis, established a proteome-based framework for assessing prognosis, and discovered neoantigens to improve breast cancer therapies [104].

Bouchal et al. used SWATH-MS to analyze 96 breast cancer tissues and cell lines. In their method, a spectral library was generated from breast cancer tissues and cell lines as a prerequisite to their breast cancer sample analysis [105]. In their findings they were able to identify about 2,800 protein groups, observed differences over the intrinsic breast cancer tumor subtypes, and discovered a greater depth of proteome variability between different breast cancer subtypes. Three differentially expressed proteins are considered to contribute strongly to improve the present breast cancer classification including **CDK1**, inositol polyphosphate 4-phosphatase type II (**INPP4B**), and **ERBB2**. Among the proteins found in the datasets, these proteins have shown the strongest correlation at the protein and transcript levels. This study was able to generate the first and only breast cancer spectral library, along with other proteomic datasets, have provided abundant sources of information for potential biomarkers research in breast cancer [105].

Rare breast cancer subtypes like mucinous carcinoma, cribriform carcinoma and tubular carcinoma respond well to few endocrine therapies. Most of the endocrine responsive subtypes usually belong to luminal A subtypes and achieve good prognosis even without neoadjuvant chemotherapy [106]. However, the following rare subtypes namely metaplastic, apocrine, adenoid cystic and medullary do not have good prognosis to endocrine therapies [107]. Although there is a small proportion of patients under these categories the chance to find specific biomarkers for these rare breast cancer subtypes is possible in the near future. Table 3, presents these current findings in summary.

#### **F. Recent breast cancer biomarkers discovery by different MS based technologies**

Mass spectrometry applied in breast cancer research has led significant findings of proteins. Blood serum of breast cancer patients was analyzed in **tandem mass spectrometry** and the study revealed protein disulfide isomerase family A, member 3 (**PDIA3**) as minimally invasive protein markers for breast cancer. PDIA3 participates in the formation of the major histocompatibility complex (MHC) class I peptide loading complex which is involved in the formation of antigens. PDIA3 was also found in MCF-7 breast cancer cell line and patients with metastatic breast cancer [108], [109] having high expression. The elevated concentration of PDIA3 is primarily due to the cell's response to stress at breast cancer state. In other findings, they discovered that such high expression of PDIA3 is correlated to TP53 gene mutation and high expression of Ki-67 antigen, which are all known to enhance proliferation of cancer cells and growth [109]–[111]. Furthermore, individuals that have TP53 germline mutation are found to be at higher risk for early-onset breast cancer [112]. A study among Malaysian women was able to use the same MS approach and found novel

proteins that are differently expressed in either stage 2 or stage 3 breast cancer cohorts only [113]. The stage

Table 3. Potential protein biomarkers for breast cancer molecular subtyping

Proteins quantified	No. of sample	Proteins differentially expressed	Significant findings	Reference
1000	6	200	Potential biomarkers for ER- BC - fascin, death-associated protein 5, Iprin- $\alpha$ 1, and $\beta$ -arrestin 1	[98]
18	18	298	Proteins involved in alterations of downstream transcription factors (TFs) in malignant tumor tissues	[99]
3500	123	10	(AIFM1), (AP1G1), (AP1M1), (CAPZB), (CMPK1), (CTNNA1), (EML4), (FTH1), GANAB, (STX12) were found on distant metastatic TNBC	[100]
9000	38	4	(CGN), (G3BP2), (PDCD4), and (OCIAD1). are potential biomarkers for the tamoxifen-susceptibility in recurrent BC	[101]
10000	40	19	Developed SVM based BC subtyping	[102]
11000	77		Proteomic analysis was comparable to mRNA BC subtyping	[103]
14000	45	9995	Identified CoTC for BC subtyping, comparably similar to PAM50 gene subtyping	[104]
2800	96	3	CDK1, (INPP4B), and ERBB2 are strongly correlated at the protein and transcript levels in BC	[105]

2 breast cancer tumor when compared to its normal adjacent tissue shows to have high expressions of the **prolyl 3-hydroxylase 1 (P3H1)**, **transmembrane emp24 domain-containing protein 10 (TMED10)**, **peptidyl-prolyl cis-trans isomerase FKBP10 (FKBP10)**, **peptidyl-prolyl cis-trans isomerase FKBP9 (FKBP9)**, **immunoglobulin superfamily containing leucine-rich repeat protein (ISLR)**, **MOB kinase activator 1A (MOB1A)**, **protein enabled homolog (ENAH)**, **collagen alpha-1(V) chain (COL5A1)**, **CAP-Gly domain-containing linker protein 1 (CLIP1)**, **protein canopy homolog 4 (COL4A1)**, **perilipin-4 (PLIN4)**, and **zinc finger CCCH domain-containing protein 18 (C3H18)**. These proteins in the stage 2 breast cancer tumor tissues are involved in many signaling pathways linked to invasion, proliferation, and migration. Protein found significantly expressed in the stage 3 breast cancer tumor includes **TAR DNA-binding protein 43 (TADBP)**, **V-type proton ATPase subunit E 1 (VATE1)**, **eukaryotic peptide chain**

**release factor subunit 1 (ETF1)**, **nucleoside diphosphate kinase 3 (NME3)**, **deoxynucleoside triphosphate triphosphohydrolase SAMHD1 (SAMH1)**, **protein SEC13 homolog (SEC13)**, **protein enabled homolog (ENAH)**, **DNA-dependent protein kinase catalytic subunit (PRKDC)**, **golgi resident protein (GCP60)**, **transmembrane glycoprotein NMB (GPNMB)**, **rho GTPase-activating protein 1 (RHG01)**, **LEM domain-containing protein 2 (LEMD2)**, **prefoldin subunit 1, coiled-coil domain-containing protein 58 (CCCD58)**, **inhibitor of nuclear factor kappa-B kinase-interacting protein (IKIP)**, **MOB kinase activator 1A (MO1A)**, and **MOB kinase activator 1B (MOB1B)**. Several among these proteins in the stage 3 breast cancer tumor tissues are found to function in supporting metastasis. Going for smaller quantities, a nanoLC-MS/MS technology was used in another study and revealed that EGF-like repeat and discoidin I-like domain-containing protein 3 (**EDIL3**) was elevated in the circulating extracellular vesicles [114]. The protein was observed to have a critical role in the integrin-FAK signaling cascade. When EDIL3 is inactivated in the MDA-MB-231 breast cancer cell lines integrin-FAK signaling pathway was also suppressed. The inactivity of EDIL3 resulted in controlling intracellular signal transduction which minimizes potential cellular invasion.

Phosphorylation of proteins is involved in the regulatory process in cells for proper functioning to occur. Phosphoproteins from biofluids are considered in many related studies as potential biomarkers for breast cancer. However, it is very challenging to find them in their native forms since they are usually denatured due to the presence of phosphatases. This is also the reason why perhaps most phosphoproteins are never detected that easy [115]. But in a recent study of Chen, et.al., they have successfully presented in their study a strategy to discover these phosphoproteins in human plasma by isolating them first from extracellular vesicles (EV) instead [115]. **SILAC** method was also employed in another phosphoproteomic study of kinase suppressor of ras-1 (**KSR1**) regulated phosphoprotein isolated from MCF-7 cell lines [116]. KSR1 phosphoprotein belongs to the RAF family of pseudokinases which is considered to play prominent roles in breast cancer growth. In the laboratory, KSR1 was observed to activate **B-Raf (BRAF) proto-oncogene** catalytic activity upon the binding of the formed protein complex between KSR1 and **mitogen-activated protein kinase (MEK)**. BRAF upon activation is directly involved in many cellular regulatory processes including cell proliferation and regulation of transcription. A mutant BRAF gene also results to a malfunctioning protein **B-raf** and has been implicated in several cancer studies including breast cancer [117]. In a study done by Jiang et al., MDA-MB-231 metastatic breast cancer cell lines were labeled by **iTRAQ** [118] to study the properties of Ras-related protein Rab-1B (**RAB1B**). Mechanistically, it was found that a down-

regulation of RAB1B in cell lines activated T $\beta$ R signaling pathway causing the expression of TGF- $\beta$  receptor 1 (**TGF $\beta$ R1**) protein levels to be elevated. Though little is known about **TGF $\beta$ R1**, one study supported its role as a tumor suppressor in the early stages of breast cancer, yet ironically it is also the same oncoprotein that may promote growth of tumor among invasive breast cancer patients [119]. In another study that also used iTRAQ was able to profile metastatic breast cancer tissues and revealed that decorin (**DCN**) and the heat shock protein 90 beta family member 1 (**HSP90B1**) as potential breast cancer biomarkers. DCN and HSP90B1 were both associated in biological pathways related to tumorigenesis promoting cell proliferation, migration, and transcription. DCN's oncogenic role was found related to tumor microenvironment showing significant interactions with EGFR and MAPK. HSP90B1 overexpression was best illustrated in a study of Huang, et al employing proteogenomics integration. They suggested that genomics alone with proteomics diagnosed by MS is not enough to show some of the phosphoprotein events that are involved in the regulatory events in the cell. Both DCN and HSP90B1 proteins are continuously being investigated at the gene and protein levels to study their roles in promoting tumor invasion and metastasis especially among TNBC patients [118]. At present, these oncoproteins are also considered indicators for poor breast cancer prognosis. The study of Lawrence et al., combined proteomics with genomic aberrations affecting protein expression [124]. They successfully discovered potential biomarkers for drug sensitivity screening based on combining the strengths of these two methods [124].

TMT can also be effective with enrichment protocols for deeper phosphoproteome analysis. Using **6-plex TMT**, Chen et al., incorporated MS approach in their drug discovery study to investigate the expression inhibition effect on the C-X-C chemokine receptor type 4 (**CXCR4**) by the **ginsenoside Rg3**, a compound extracted from Panax ginseng [120]. MDA-MB-231 cell line treated with ginsenoside Rg3 peptides were labeled with 6-plex TMT. The mixture was divided into fractions and each fraction subjected to TiO<sub>2</sub>-based phosphopeptide enrichment before the LC-MS/MS analysis. The protein analysis showed that the Rg3 protein were found to display an inhibitory effect against CXCR4 and targets the anti-inflammatory nuclear factor-kB signaling pathway [121]. These findings suggested that ginsenoside Rg3 have chemopreventive properties that may control metastasis of breast cancer cells.

Staging breast cancer is another interesting aspect for proteomic study. In the lab of Lobo et. al., they investigated using **DIA label-free MS method** to analyze protein expression from patient blood samples at different breast cancer stages. Three specific proteins were found to have varied expressions at different stages of breast cancer. These proteins are **clusterin**, **apolipoprotein A-II (APOA2)**, and **apolipoprotein C-**

**III (APOC3)**. These proteins play important roles in alterations related to protein glycosylation, progression and invasiveness of breast cancer [122]. Aside from tissues and blood samples, some studies made use of urine to detect biomarkers in label-free MS analysis. In the study of Beretov et al., successfully identified 13 proteins from urine for its potential utility to detect preinvasive breast cancer stage to metastatic breast cancer stage among DCIS patients [123]. The novel proteins they discovered from the tissue samples include **cytosolic non-specific dipeptidase (PEPA)**, **multimerin-2 (MMRN2)**, **neuronal growth regulator 1 (NEGR1)**, **leucine-rich repeat-containing protein 36 (LRC36)**, **microtubule-associated serine/threonine-protein kinase 4 (MAST4)**, **keratin, type I cytoskeletal 10 (K1C10)**, **uncharacterized protein C9orf131 (C1131)**, **uncharacterized protein C4orf14 (CD014)**, **filaggrin**, **dynein heavy chain 8, axonemal (DYH8)**, **hemoglobin subunit alpha (HBA)**, **AGRIN**, and **fibrinogen alpha chain (FIBA)**. From this list, they tested their hypothesis only against MAST4 to validate the protein profile in both tissue and urine samples. The result of their study was able to detect the protein biomarkers in both body derivatives.

These are just some of the many on-going research that are exploring the application of various MS techniques and strategies that will lead to the discovery of novel protein tumor biomarkers for breast cancer. It is important to note that at this rate, mass spectrometry-based research is going very fast in integrating label-free MS methods together with library free computational tools approach to drive its workable utility in the clinical setting.

### CONCLUSIONS AND FUTURE DIRECTIONS

The high throughput proteomic methods applied in cancer research have been evolving dramatically at an increasing pace. The majority of these protein profiling studies exhibit promising diagnostic, prognostic or predictive values in controlling and mitigating breast cancer. More study is in place to unravel breast cancer's nature, molecular features, and tumor biology to explain its uncontrolled growth and development. By using mass spectrometry and its computational tools for identification and quantitation it would not be long that soon this method would completely be used for clinical applications.

At present, breast cancer intrinsic protein biomarkers are routinely measured primarily by IHC using antibody-based techniques or by genomic-based test. Although both mRNA extraction kits require highly specific antibodies, genetic tests are also costly and are often effective only to patients expressing hormone receptors. IHC may also see some issues wherein the immunoreactivity can be compromised because of post-translational modifications of the proteins. Furthermore, without quantifiable biomarkers like in the case of the TNBC subtype, it would be challenging to come up with an exact regimen for medical treatment.

Today, mass spectrometry has seriously been reengineered to full capability to analyze down to nanoparticle size samples at high throughput. The instrument and software developed for database search and analysis has remarkably enabled us to investigate in-depth breast cancer more effectively. It was also noted earlier, that there are different proteins involved in the development of the mammary gland, but very few are published about the proteomic landscape of breast cancer patients from young, premenopausal, to postmenopausal adults. Also, staging has been noted to be well studied even in mass spectrometry, but the proteomic profile accounting for breast cancer grades has not been well studied using this approach. These missing links in the biology of breast cancer may soon unfold especially with updated databases uploaded in the human proteome atlas (HPA) [124]. This open-access online platform is an advantage for researchers to explore the properties and functions of every protein. The HPA database includes a list of about 17000 unique proteins, 26000 antibodies, protein cellular locations, functions, cellular processes, and disease associated proteins including those involved in breast cancer [124]. The information in this database was mostly acquired from mass spectrometry-based proteomics data acquisition, other omics-based technologies and a variety of immuno-based methods. With new information that MS technology can derived, more novel proteins are to be found and mapped to networks and pathways associated to breast cancer development and use them for effective drug design and therapy.

## REFERENCES

- [1] WHO, *Who Report on Cancer*. 2020.
- [2] NCI, "NCI Center to Reduce Cancer Health Disparities," 2019. [Online]. Available: <https://www.cancer.gov/about-nci/organization/crhd>. [Accessed: 28-May-2019].
- [3] A.Cooper, *On the anatomy of the breast*. Jefferson, 1840.
- [4] M. L.Marinovich *et al.*, "Agreement between MRI and pathologic breast tumor size after neoadjuvant chemotherapy, and comparison with alternative tests: individual patient data meta-analysis," *BMC Cancer*, vol. 15, p. 662, Oct.2015.
- [5] M.Mayo Clinic, "Breast Cancer," 2019, 2019. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475>.
- [6] C. J.Watson andW. T.Khaled, "Mammary development in the embryo and adult: a journey of morphogenesis and commitment," *Development*, vol. 135, no. 6, pp. 995 LP – 1003, Mar.2008.
- [7] J. H.Krege *et al.*, "Generation and reproductive phenotypes of mice lacking estrogen receptor beta," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 26, pp. 15677–15682, Dec.1998.
- [8] W.Ruan andD. L.Kleinberg, "Insulin-Like Growth Factor I Is Essential for Terminal End Bud Formation and Ductal Morphogenesis during Mammary Development1," *Endocrinology*, vol. 140, no. 11, pp. 5075–5081, Nov.1999.
- [9] J.McBryan, J.Howlin, S.Napoletano, andF.Martin, "Amphiregulin: Role in Mammary Gland Development and Breast Cancer," *J. Mammary Gland Biol. Neoplasia*, vol. 13, pp. 159–169, Jul.2008.
- [10] S. R.Oakes, H. N.Hilton, andC. J.Ormandy, "The alveolar switch: coordinating the proliferative cues and cell fate decisions that drive the formation of lobuloalveoli from ductal epithelium," *Breast Cancer Res.*, vol. 8, no. 2, p. 207, 2006.
- [11] T. R.Milanese *et al.*, "Age-related lobular involution and risk of breast cancer.," *J. Natl. Cancer Inst.*, vol. 98, no. 22, pp. 1600–1607, Nov.2006.
- [12] V.Sangwan andM.Park, "Receptor tyrosine kinases: role in cancer progression," *Curr. Oncol.*, vol. 13, no. 5, pp. 191–193, Oct.2006.
- [13] S. P.Chiang, G. S.Karagiannis, J. S.Condeelis, andJ. E.Segall, "Abstract 1844: Investigating the role of amphiregulin in breast cancer," *Cancer Res.*, vol. 77, no. 13 Supplement, pp. 1844 LP – 1844, Jul.2017.
- [14] S.Lee *et al.*, "Alterations of gene expression in the development of early hyperplastic precursors of breast cancer.," *Am. J. Pathol.*, vol. 171, no. 1, pp. 252–262, Jul.2007.
- [15] S.Mao *et al.*, "Abstract 5114: The role of amphiregulin in mammary gland development and breast cancer," *Cancer Res.*, vol. 78, no. 13 Supplement, pp. 5114 LP – 5114, Jul.2018.
- [16] A.Baillo, C.Giroux, andS. P.Ethier, "Knock-down of amphiregulin inhibits cellular invasion in inflammatory breast cancer.," *J. Cell. Physiol.*, vol. 226, no. 10, pp. 2691–2701, Oct.2011.
- [17] S.Rao, S. J. F.Cronin, V.Sigl, andJ. M.Penninger, "RANKL and RANK: From Mammalian Physiology to Cancer Treatment," *Trends Cell Biol.*, vol. 28, no. 3, pp. 213–223, 2018.
- [18] E.Gonzalez-Suarez *et al.*, "RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis.," *Nature*, vol. 468, no. 7320, pp. 103–107, Nov.2010.
- [19] D.Schramek *et al.*, "Osteoclast differentiation factor RANKL controls development of progestin-driven mammary cancer.," *Nature*, vol. 468, no. 7320, pp. 98–102, Nov.2010.
- [20] D. C.Radisky andL. C.Hartmann, "Mammary involution and breast cancer risk: transgenic models and clinical studies," *J. Mammary Gland Biol. Neoplasia*, vol. 14, no. 2, pp. 181–191, Jun.2009.
- [21] R. W. E.Clarkson andC. J.Watson, "Microarray analysis of the involution switch.," *J. Mammary Gland Biol. Neoplasia*, vol. 8, no. 3, pp. 309–319, Jul.2003.
- [22] L. T.Bemis andP.Schedin, "Reproductive state of rat mammary gland stroma modulates human breast cancer cell migration and invasion.," *Cancer Res.*, vol. 60, no. 13, pp. 3414–3418, Jul.2000.

- [23] M.Xue *et al.*, "Regulation of estrogen signaling and breast cancer proliferation by an ubiquitin ligase TRIM56," *Oncogenesis*, vol. 8, no. 5, p. 30, 2019.
- [24] J.Wang and B.Xu, "Targeted therapeutic options and future perspectives for HER2-positive breast cancer," *Signal Transduct. Target. Ther.*, vol. 4, no. 1, p. 34, 2019.
- [25] L.Xia *et al.*, "Role of the NF $\kappa$ B-signaling pathway in cancer," *Onco. Targets. Ther.*, vol. 11, pp. 2063–2073, Apr.2018.
- [26] T. H. B.Gomig *et al.*, "Quantitative label-free mass spectrometry using contralateral and adjacent breast tissues reveal differentially expressed proteins and their predicted impacts on pathways and cellular functions in breast cancer," *J. Proteomics*, vol. 199, no. August 2018, pp. 1–14, 2019.
- [27] E.Presti, Daniele and Quaquarini, "The PI3K/AKT/ mTOR and CDK4/6 Pathways in Endocrine Resistant HR+/HER2- Metastatic Breast Cancer: Biological Mechanisms and New Treatments," *Cancers (Basel)*, vol. 11, no. 9, pp. 1–20, 2019.
- [28] B. M.Krishna *et al.*, "Notch signaling in breast cancer: From pathway analysis to therapy," *Cancer Lett.*, vol. 461, pp. 123–131, 2019.
- [29] M. A.Gillette *et al.*, "Proteogenomics connects somatic mutations to signalling in breast cancer," *Nature*, pp. 1–19, 2016.
- [30] H. J.Burstein, "The distinctive nature of HER2-positive breast cancers.," *N. Engl. J. Med.*, vol. 353, no. 16, pp. 1652–1654, Oct.2005.
- [31] NIH, "Breast Cancer," 2019. [Online]. Available: <https://ghr.nlm.nih.gov/condition/breast-cancer#resources>.
- [32] K. A.Metcalf *et al.*, "The risk of breast cancer in BRCA1 and BRCA2 mutation carriers without a first-degree relative with breast cancer.," *Clin. Genet.*, vol. 93, no. 5, pp. 1063–1068, May2018.
- [33] Z.Mitri, T.Constantine, and R.O'Regan, "The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy," *Chemother. Res. Pract.*, vol. 2012, p. 743193, 2012.
- [34] M. C.Figueroa-Magalhães, D.Jelovac, R.Connolly, and A. C.Wolff, "Treatment of HER2-positive breast cancer," *Breast*, vol. 23, no. 2, pp. 128–136, Apr.2014.
- [35] M.Dowsett *et al.*, "Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group," *J. Natl. Cancer Inst.*, vol. 103, no. 22, pp. 1656–1664, Nov.2011.
- [36] M.Abubakar *et al.*, "Combined quantitative measures of ER , PR , HER2 , and KI67 provide more prognostic information than categorical combinations in luminal breast cancer," *Mod. Pathol.*, pp. 1244–1256, 2019.
- [37] F. J.Candido *et al.*, "An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation," pp. 1–13, 2017.
- [38] A.Vaishnavi, A. T.Le, and R. C.Doebele, "TRKing down an old oncogene in a new era of targeted therapy.," *Cancer Discov.*, vol. 5, no. 1, pp. 25–34, Jan.2015.
- [39] J. J.Lee, K.Loh, and Y.-S.Yap, "PI3K/Akt/mTOR inhibitors in breast cancer," *Cancer Biol. Med.*, vol. 12, no. 4, pp. 342–354, Dec.2015.
- [40] O.Willis *et al.*, "PIK3CA gene aberrancy and role in targeted therapy of solid malignancies," *Cancer Gene Ther.*, 2020.
- [41] Y. F.Hou, M.F., Chen, Y.L., Tseng, T.F., Lin, C.M., Chen, M.S., Huang, C.J., Huang, Y.S., Hsieh, J.S., Huang, T.J., Jong, S.B., Huang, "Evaluation of serum CA27.29, CA15-3 and CEA in patients with breast cancer," *Kaohsiung J. Med. Sci.*, 1999.
- [42] C.VanPoznak, L. N.Harris, and M. R.Somerfield, "Use of Biomarkers to Guide Decisions on Systemic Therapy for Women With Metastatic Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline," *J. Oncol. Pract.*, vol. 11, no. 6, pp. 514–516, Jul.2015.
- [43] A. M.Kabel, "Tumor markers of breast cancer: New prospectives," *J. Oncol. Sci.*, vol. 3, no. 1, pp. 5–11, 2017.
- [44] X.Zhe, M. L.Chen, and R. D.Bonfil, "Circulating tumor cells: finding the needle in the haystack," *Am. J. Cancer Res.*, vol. 1, no. 6, pp. 740–751, 2011.
- [45] W. A.Osta *et al.*, "EpCAM Is Overexpressed in Breast Cancer and Is a Potential Target for Breast Cancer Gene Therapy," *Cancer Res.*, vol. 64, no. 16, pp. 5818 LP – 5824, Aug.2004.
- [46] S.Cotterill, "Cancer Genetics Web," 2019. [Online]. Available: <http://www.cancer-genetics.org/PTPRC.htm>. [Accessed: 10-Jul-2020].
- [47] E.Jarasch, R. A. Y. B.Nagle, M.Kaufmann, C.Maurer, and W. J.Bocker, "Differential Diagnosis of Benign Epithelial Proliferations and Carcinomas of the Breast Using Antibodies to Cytokeratins," 1987.
- [48] Y.Jing, J.Zhang, S.Waxman, and R.Mira-y-Lopez, "Upregulation of cytokeratins 8 and 18 in human breast cancer T47D cells is retinoid-specific and retinoic acid receptor-dependent.," *Differentiation.*, vol. 60, no. 2, pp. 109–117, May1996.
- [49] M. B.Lustberg *et al.*, "Heterogeneous atypical cell populations are present in blood of metastatic breast cancer patients.," *Breast Cancer Res.*, vol. 16, no. 2, p. R23, Mar.2014.
- [50] C. P .Leo, C.Leo, and T. D.Szucs, "Breast cancer drug approvals by the US FDA from 1949 to 2018," *Nat. Rev. Drug Discov.*, vol. 19, no. 1, p. 11, 2020.
- [51] A.Garcia-Diaz *et al.*, "Interferon Receptor Signaling Pathways Regulating PD-L1 and PD-L2 Expression.," *Cell Rep.*, vol. 19, no. 6, pp. 1189–1201, May2017.
- [52] Y.Han, D.Liu, and L.Li, "PD-1/PD-L1 pathway: current researches in cancer," *Am. J. Cancer Res.*, vol. 10, no. 3, pp. 727–742, Mar.2020.
- [53] F.Schütz, S.Stefanovic, L.Mayer, A.VonAu, C.Domschke, and C.Sohn, "PD-1/PD-L1 Pathway in Breast Cancer," *Oncol. Res. Treat.*, vol. 40, no. 5, pp. 294–297, 2017.
- [54] A.Illiano, G.Pinto, C.Melchiorre, A.Carpentieri, V.Faraco, and A.Amoresano, "Protein Glycosylation Investigated by Mass Spectrometry: An Overview," *Cells*, vol. 9, no. 9, pp. 1–22, 2020.

- [55] K. A. Resing and N. G. Ahn, "Proteomics strategies for protein identification," vol. 579, pp. 885–889, 2005.
- [56] G. Büyükköroğlu, D. D. Dora, F. Özdemir, and C. Hizel, "Chapter 15 - Techniques for Protein Analysis," D. Barh and V. B. T.-O. T. and B.-E. Azevedo, Eds. Academic Press, 2018, pp. 317–351.
- [57] T. J. Griffin and R. Aebersold, "Advances in Proteome Analysis by Mass Spectrometry," *J. Biol. Chem.*, vol. 276, no. 49, pp. 45497–45500, 2001.
- [58] S. B. Nukala, G. Baron, G. Aldini, M. Carini, and A. D'Amato, "Mass Spectrometry-based Label-free Quantitative Proteomics To Study the Effect of 3PO Drug at Cellular Level.," *ACS Med. Chem. Lett.*, vol. 10, no. 4, pp. 577–583, Apr. 2019.
- [59] M. W. Duncan and S. W. Hunsucker, "Proteomics as a tool for clinically relevant biomarker discovery and validation.," *Exp. Biol. Med. (Maywood)*, vol. 230, no. 11, pp. 808–817, Dec. 2005.
- [60] F. M. White, "Quantitative phosphoproteomic analysis of signaling network dynamics.," *Curr. Opin. Biotechnol.*, vol. 19, no. 4, pp. 404–409, Aug. 2008.
- [61] A. P. Frei *et al.*, "Direct identification of ligand-receptor interactions on living cells and tissues.," *Nat. Biotechnol.*, vol. 30, no. 10, pp. 997–1001, Oct. 2012.
- [62] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.
- [63] F. Desiere *et al.*, "The PeptideAtlas project.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. 655–658, 2006.
- [64] G. Rosenberger *et al.*, "A repository of assays to quantify 10,000 human proteins by SWATH-MS," *Sci. Data*, vol. 1, no. 1, p. 140031, 2014.
- [65] P. Palmowski *et al.*, "The Generation of a Comprehensive Spectral Library for the Analysis of the Guinea Pig Proteome by SWATH-MS," *Proteomics*, vol. 19, no. 15, pp. 1–5, 2019.
- [66] Y. Lin, "In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics," *Nat. Commun.*, pp. 1–11, 2020.
- [67] E. Hoedt, G. Zhang, and T. A. Neubert, "Stable isotope labeling by amino acids in cell culture (SILAC) for quantitative proteomics.," *Adv. Exp. Med. Biol.*, vol. 806, pp. 93–106, 2014.
- [68] S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid, "Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.," *Proteomics*, vol. 7, no. 3, pp. 340–350, Feb. 2007.
- [69] J. Muntel *et al.*, "Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows with Fixed Instrument Time," 2019.
- [70] L. R. Zieske, "A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies.," *J. Exp. Bot.*, vol. 57, no. 7, pp. 1501–1508, 2006.
- [71] L. Zhang and J. E. Elias, "Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry BT - Proteomics: Methods and Protocols," L. Comai, J. E. Katz, and P. Mallick, Eds. New York, NY: Springer New York, 2017, pp. 185–198.
- [72] D. Virág *et al.*, "Current Trends in the Analysis of Post-translational Modifications," *Chromatographia*, vol. 83, no. 1, pp. 1–10, 2020.
- [73] B. Manadas, V. M. Mendes, J. English, and M. J. Dunn, "Peptide fractionation in proteomics approaches.," *Expert Rev. Proteomics*, vol. 7, no. 5, pp. 655–663, Oct. 2010.
- [74] E. R. Sauter, "Reliable Biomarkers to Identify New and Recurrent Cancer," *Eur. J. breast Heal.*, vol. 13, no. 4, pp. 162–167, Oct. 2017.
- [75] E. M. Woo, D. Fenyo, B. H. Kwok, H. Funabiki, and B. T. Chait, "Efficient Identification of Phosphorylation by Mass Spectrometric Phosphopeptide Fingerprinting," *Anal. Chem.*, vol. 80, no. 7, pp. 2419–2425, Apr. 2008.
- [76] S. A. Beausoleil *et al.*, "Large-scale characterization of HeLa cell nuclear phosphoproteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 33, pp. 12130–12135, 2004.
- [77] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. C. Collins, and R. Aebersold, "Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial," *Mol. Syst. Biol.*, vol. 14, no. 8, p. e8126, Aug. 2018.
- [78] Y. S. Ting *et al.*, "Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data.," *Mol. Cell. Proteomics*, vol. 14, no. 9, pp. 2301–2307, Sep. 2015.
- [79] S. Tyanova, T. Temu, and J. Cox, "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics," *Nat. Protoc.*, vol. 11, p. 2301, Oct. 2016.
- [80] O. Bernhardt *et al.*, *Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data*. 2014.
- [81] R. C. Poulos *et al.*, "Strategies to enable large-scale proteomics for reproducible research," *Nat. Commun.*, vol. 11, no. 1, pp. 1–13, 2020.
- [82] Institute for Systems Biology, "Spectrum Library Central at PeptideAtlas," 2015. [Online]. Available: <http://www.peptideatlas.org/specclib/>. [Accessed: 08-Oct-2019].
- [83] H. L. Röst *et al.*, "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.," *Nature biotechnology*, vol. 32, no. 3, United States, pp. 219–223, Mar. 2014.
- [84] J. Wang *et al.*, "MSPLIT-DIA: sensitive peptide identification for data-independent acquisition.," *Nature methods*, vol. 12, no. 12, pp. 1106–1108, Dec. 2015.
- [85] Y. Yang, X. Liu, C. Shen, Y. Lin, P. Yang, and L. Qiao, "In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics," *Nat. Commun.*, vol. 11, no. 1, p. 146, Jan. 2020.
- [86] S. Gessulat *et al.*, "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nat. Methods*, vol. 16, no. 6, pp. 509–518, 2019.
- [87] N. H. Tran *et al.*, "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry.," *Nat. Methods*, vol. 16, no. 1, pp. 63–66, Jan. 2019.

- [88] A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev, and P. A. Pevzner, "De novo peptide sequencing and identification with precision mass spectrometry," *J. Proteome Res.*, vol. 6, no. 1, pp. 114–123, Jan. 2007.
- [89] R. S. Johnson and J. A. Taylor, "Searching sequence databases via De novo peptide sequencing by tandem mass spectrometry," *Mol. Biotechnol.*, vol. 22, no. 3, pp. 301–315, 2002.
- [90] B. Ma *et al.*, "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 17, no. 20, pp. 2337–2342, 2003.
- [91] Y. Li *et al.*, "Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files," *Nature methods*, vol. 12, no. 12, United States, pp. 1105–1106, Dec-2015.
- [92] Y. S. Ting *et al.*, "PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data," *Nat. Methods*, vol. 14, no. 9, pp. 903–908, Sep. 2017.
- [93] C.-C. Tsou *et al.*, "DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics," *Nat. Methods*, vol. 12, no. 3, pp. 258–64, 7 p following 264, Mar. 2015.
- [94] B. C. Searle *et al.*, "Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry," *Nat. Commun.*, vol. 9, no. 1, p. 5128, 2018.
- [95] P. Pandit, R. Patil, V. Palwe, S. Gandhe, R. Patil, and R. Nagarkar, "Prevalence of Molecular Subtypes of Breast Cancer: A Single Institutional Experience of 2062 Patients," *Eur. J. breast Heal.*, vol. 16, no. 1, pp. 39–43, Nov. 2019.
- [96] C. K. Y. Ng, A. M. Schultheis, F.-C. Bidard, B. Weigelt, and J. S. Reis-Filho, "Breast cancer genomics from microarrays to massively parallel sequencing: paradigms and new insights," *J. Natl. Cancer Inst.*, vol. 107, no. 5, Feb. 2015.
- [97] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.
- [98] K. Rezaul *et al.*, "Differential protein expression profiles in estrogen receptor-positive and -negative breast cancer tissues using label-free quantitative proteomics," *Genes and Cancer*, vol. 1, no. 3, pp. 251–271, 2010.
- [99] B. L. K. S. Cha, M. B. Imielinski, T. Rejtár, E. A. Richardson, D. Thakur, D. C. Sgroi, "In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology *Mol Cell Proteomics*, 9 (2010), pp. 2529–2544," 2010.
- [100] N. Q. Liu *et al.*, "Comparative proteome analysis revealing an 11-protein signature for aggressive triple-negative breast cancer," *J. Natl. Cancer Inst.*, vol. 106, no. 2, 2014.
- [101] A. U. T. De Marchi, N. Q. Liu, C. Stingl, M. A. Timmermans, M. Smid, M. P. Look, M. Tjoa, R. B. Braakman, M. Opdam, S. C. Linn, F. C. Sweep, P. N. Span, M. Kliffen, T. M. Luider, J. A. Foekens, J. W. Martens, "4-protein signature predicting tamoxifen treatment outcome in recurrent breast cancer *Mol Oncol*, 10 (2016), pp. 24–39," 2016.
- [102] S. Tyanova, R. Albrechtsen, P. Kronqvist, J. Cox, M. Mann, and T. Geiger, "Proteomic maps of breast cancer subtypes," *Nat. Commun.*, vol. 7, no. 1, p. 10259, 2016.
- [103] P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, "Proteogenomics connects somatic mutations to signalling in breast cancer *Nature*, 534 (2016), pp. 55–62," 2016.
- [104] H. J. Johansson *et al.*, "Breast cancer quantitative proteome and proteogenomic landscape," *Nat. Commun.*, vol. 10, no. 1, p. 1600, 2019.
- [105] P. Bouchal *et al.*, "Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry," *Cell Rep.*, vol. 28, no. 3, pp. 832–843.e7, Jul. 2019.
- [106] T. Nagao, T. Kinoshita, T. Hojo, H. Tsuda, K. Tamura, and Y. Fujiwara, "The differences in the histological types of breast cancer and the response to neoadjuvant chemotherapy: the relationship between the outcome and the clinicopathological characteristics," *Breast*, vol. 21, no. 3, pp. 289–295, Jun. 2012.
- [107] M. V. Dieci, E. Orvieto, M. Dominici, P. Conte, and V. Guarneri, "Rare Breast Cancer Subtypes: Histological, Molecular, and Clinical Peculiarities," pp. 805–813, 2014.
- [108] H. Zou *et al.*, "P4HB and PDIA3 are associated with tumor progression and therapeutic outcome of diffuse gliomas," *Oncol. Rep.*, vol. 39, no. 2, pp. 501–510, Feb. 2018.
- [109] M. Song *et al.*, "Proteomic Analysis of Breast Cancer Tissues to Identify Biomarker Candidates by Gel-Assisted Digestion and Label-Free Quantification Methods Using LC-MS / MS," vol. 35, no. 10, pp. 1839–1847, 2012.
- [110] L. T. Li, G. Jiang, Q. Chen, and J. N. Zheng, "Predic Ki67 is a promising molecular target in the diagnosis of cancer (Review)," *Mol. Med. Rep.*, vol. 11, no. 3, pp. 1566–1572, 2015.
- [111] F. Mantovani, L. Collavin, and G. Del Sal, "Mutant p53 as a guardian of the cancer cell," *Cell Death Differ.*, vol. 26, no. 2, pp. 199–212, 2019.
- [112] A. D. Sorrell, C. R. Espenschied, J. O. Culver, and J. N. Weitzel, "Tumor protein p53 (TP53) testing and Li-Fraumeni syndrome: current status of clinical applications and future directions," *Mol. Diagn. Ther.*, vol. 17, no. 1, pp. 31–47, Feb. 2013.
- [113] A. S. Al-Wajeeh *et al.*, "Comparative proteomic analysis of different stages of breast cancer tissues using ultra high performance liquid chromatography tandem mass spectrometer," *PLoS One*, vol. 15, no. 1, pp. e0227404–e0227404, Jan. 2020.
- [114] J.-E. Lee *et al.*, "Identification of EDIL3 on extracellular vesicles involved in breast cancer cell invasion," *J. Proteomics*, vol. 131, pp. 17–28, 2016.
- [115] I.-H. Chen *et al.*, "Phosphoproteins in extracellular vesicles as candidate markers for breast cancer," *Proc.*



- Natl. Acad. Sci.*, vol. 114, no. 12, pp. 3175 LP – 3180, Mar.2017.
- [116] H.Zhang *et al.*, “SILAC-based phosphoproteomics reveals an inhibitory role of KSR1 in p53 transcriptional activity via modulation of DBC1,” *Br. J. Cancer*, vol. 109, no. 10, pp. 2675–2684, Nov.2013.
- [117] D.Liu andK.Zhou, “BRAF/MEK Pathway is Associated With Breast Cancer in ER-dependent Mode and Improves ER Status-based Cancer Recurrence Prediction,” *Clin. Breast Cancer*, vol. 20, no. 1, pp. 41-50.e8, 2020.
- [118] H.-L.Jiang *et al.*, “Loss of RAB1B promotes triple-negative breast cancer metastasis by activating TGF-beta/SMAD signaling,” *Oncotarget*, vol. 6, no. 18, pp. 16352–16365, Jun.2015.
- [119] B.Tang *et al.*, “TGF-beta switches from tumor suppressor to prometastatic factor in a model of breast cancer progression,” *J. Clin. Invest.*, vol. 112, no. 7, pp. 1116–1124, Oct.2003.
- [120] X. C. L.Qian, H.Jiang, andJ.Chen, “Ginsenoside Rg3 inhibits CXCR 4 expression and related migrations in a breast cancer cell line,” pp. 519–523, 2011.
- [121] M.Zou, J.Wang, J.Gao, H.Han, andY.Fang, “Phosphoproteomic analysis of the antitumor effects of ginsenoside Rg3 in human breast cancer cells,” *Oncol. Lett.*, vol. 15, no. 3, pp. 2889–2898, Mar.2018.
- [122] M. D. P.Lobo, F. B. M. B.Moreno, G. H. M. F.Souza, S. M. M. L.Verde, R. de A.Moreira, andA. C. de O.Monteiro-Moreira, “Label-free proteome analysis of plasma from patients with breast cancer: Stage-specific protein expression,” *Front. Oncol.*, vol. 7, no. FEB, pp. 1–12, 2017.
- [123] J.Beretov, V. C.Wasinger, E. K. A.Millar, P.Schwartz, P. H.Graham, andY.Li, “Proteomic analysis of urine to identify breast cancer biomarker candidates using a label-free LC-MS/MS approach,” *PLoS One*, vol. 10, no. 11, 2015.
- [124] “Human Proteome Atlas,” 2020. [Online]. Available: <https://www.proteinatlas.org/>.

# Identification of Class I HLA Alleles in Anonymized Cell Therapy Specimens through Real-Time PCR with Melt-Curve Analysis

Joanne Jennifer E. Tan<sup>1</sup>, Maria Teresa A. Barzaga<sup>2</sup> and Francisco M. Heralde III<sup>2,3</sup>

<sup>1</sup>College of Medicine, University of the Philippines Manila

<sup>2</sup>Molecular Diagnostics and Cellular Therapeutics Laboratory, Lung Center of the Philippines

<sup>3</sup>Department of Biochemistry and Molecular Biology, College of Medicine, University of the Philippines Manila

## Email address:

jetan1@up.edu.ph

## To cite:

Tan, JJE; Barzaga, MTA; Heralde, FIIM.2020. Identification of Class I HLA Alleles in Anonymized Cell Therapy Specimens through Real-Time PCR with Melt-Curve Analysis. PJBMB. Vol. 1, No. 1, 2020, pp. 29-36. doi: 10.5555/pjbmb.ph.2020.01.01.29

Received: 09 19, 2019; Accepted: 11 18, 2020; Published: 12 06, 2020

**Abstract:** Accurate human leukocyte antigen (HLA) typing is crucial for allogeneic dendritic cell anti-cancer vaccination, where at least a seven out of eight HLA match to the medium resolution allele level is required. Molecular methods can provide medium to high resolution typing but are expensive and time-consuming. A modified method to facilitate faster and more efficient medium resolution identity matching for some common alleles in Filipinos was developed using real-time PCR with melt-curve analysis. The most common Filipino HLA Class I alleles identified from available databases were: A\*02, A\*24, B\*15 and C\*07. Primers specific to these alleles were designed. DNA were extracted from 17 de-identified unstimulated stem cell specimens from the Molecular Diagnostics and Cellular Therapeutics Laboratory of the Lung Center of the Philippines (MDCTL-LCP). Comparison of the melt-curve profile was able to determine some degree of HLA allele identity. Distinct alleles from the samples that were identified through sequencing were A\*03:01 and C\*07:32. Higher resolution typing was not possible for some alleles due to highly similar sequences in the amplified region. The study thus explored the use of real-time PCR and melting curve analysis in describing some common HLA Class I alleles in Filipinos. Further replication and more specific primers will be needed to establish the melting curve profiles for HLA identification use.

**Keywords:** allogeneic dendritic cell vaccine; HLA Class I typing; real-time PCR with melt-curve analysis; A\*03:01 and C\*07:32 alleles

## 1. INTRODUCTION

Human Leukocyte Antigens (HLA) are cell surface proteins that present antigenic peptides to generate immune defense reactions (Choo, 2007; Cruz, 2017). The HLA loci are the most polymorphic genes present in the whole genome, which ensure that few individuals are highly similar, and that the population is well-equipped to deal with all types of immune attacks (Choo, 2007). There are two classes of HLA based on their structure and function: HLA Class I and Class II. Class I molecules, the most common ones being HLA-A, -B, and -C, are expressed on virtually all nucleated cells, whereas class II molecules, which include HLA-

DR, -DQ and -DP, are only expressed on "immune competent" cells, such as dendritic cells, B lymphocytes, and macrophages (Allard et al., 2014; Leddon et al., 2010).

HLA typing is primarily performed to prevent graft rejection and graft vs. host disease, which can happen when HLA types do not match during cell or tissue transplantation (Choi et al. 2009). There are two main categories of HLA typing: serological-based and molecular-based (Paunic et al., 2012). HLA typing by serology is the oldest and most common method in routine clinical setting (Erich et al., 2012). However, serology may be unreliable and only provides low

resolution typing, which will give a limited detection of HLA polymorphism (Erlich et al., 2012). With the development of molecular techniques, medium to high resolution methods of HLA typing are now available (Paunic et al., 2012). Molecular techniques, or DNA-based histocompatibility testing, utilize the polymerase chain reaction (PCR) which is the general method used to amplify specific regions of DNA.

The most common PCR-based HLA typing techniques include sequence-specific oligonucleotide probes (SSO), sequence-specific primers (SSP), and sequence-based typing (SBT) (Perng, et al., 2012). PCR-SSO is a relatively a high-throughput and inexpensive method, and is usually used for large-scale, low-resolution HLA allele analysis. SSP provides medium to high resolution and is typically used on samples that have failed to be analyzed by SSO, since it is more expensive and not ideal for many samples. SBT has the highest resolution and is the only way to directly sequence and identify new alleles, although it is also the most expensive. Real-time PCR is now also being used in HLA typing since it involves minimal hands-on time and thus subject to less human error, is less expensive compared to other methods, and does not require post-PCR processing, therefore reducing the risk of contamination (Gersuk & Nepom, 2006). SBT is regarded as the gold standard in HLA typing (Perng, et al., 2012).

Anti-cancer vaccination using peptide-pulsed dendritic cells depends on the interaction of the HLA molecule and the corresponding epitope. Accurate HLA-typing is therefore very crucial for successful anti-cancer vaccination, especially in the case of allogeneic vaccination for immunocompromised patients, where at least a seven out of eight HLA match to the medium resolution allele level is required. Furthermore, characterization and classification of HLA molecules into superfamilies or supertypes in terms of peptide-binding specificities is valuable for the development of anti-cancer vaccines. Previous studies have used bioinformatics to perform hierarchical clustering and principal component analysis to classify HLA class I and class II alleles into such supertypes (Doytchinova et al., 2004). In this study, a method to facilitate faster and more efficient medium resolution identity matching for some common alleles in Filipinos using real-time PCR with melt-curve analysis was explored. Using available databases for primer design, DNA from de-identified unstimulated stem cell specimens from the Molecular Diagnostics and Cellular Therapeutics Laboratory of the Lung Center of the Philippines (MDCTL-LCP) were evaluated thru real-time PCR assay with melt-curve analysis and HLA identification was validated through sequencing.

## 2. METHODOLOGY

### 2.1 Samples and DNA extraction

Previously extracted DNA (following DNeasy Blood and Tissue Extraction Kit protocol) from 17 de-identified unstimulated stem cell specimens were obtained from the MDCTL-LCP. The specimens were derived from stored samples of patients under the stem cell therapeutics program of the Lung Center of the Philippines, who accomplished informed consent forms (ICF) specifying that their blood samples, particularly unstimulated stem cells (USCs) and dendritic cells, will be collected and separated following established clinical protocols for collection from the bone marrow, adipose tissue or through leukapheresis, and stored for an indefinite amount of time in a cryolocator containing liquid nitrogen. Access to the samples are limited to the physicians and members of the MDCTL-LCP project team, with no information regarding the identity of the patient to be disclosed. The ICF includes use of the specimens for the conduct of research that are either basic or applied, with the goal of improving LCP's molecular diagnostics and therapeutics program.

### 2.2 Primer Design

Selection of supertypes. The most common alleles for the Filipino population were obtained from the dbMHC database (<https://www.ncbi.nlm.nih.gov/gv/mhc/>). Their corresponding supertypes were then identified using the classification system of Doytchinova et al. (2004).

Primer design. Primers specific to the chosen alleles were designed using Primer-BLAST with standard settings (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>) (Ye et al., 2012). Primers with amplicon lengths around 200 bp (range: 178-220) were selected arbitrarily for its intended use for real-time PCR, as longer lengths do not amplify as efficiently (Smith and Osborn, 2009). The primers were made to amplify exon 2 or 3 of the selected HLA allele, since these are the regions known to contain the most polymorphisms (total length of alleles ~1000 bp). Alleles were aligned using the IMGT/ HLA database (<https://www.ebi.ac.uk/ipd/imgt/hla/index.html>) (Li et al., 2015). The designed primer sequences and amplicon lengths are found in **Table 2**. Primers were also cross-checked through the NCBI database to ensure that there it will not amplify other regions in the genome.

### 2.3 Optimization of the Real-Time PCR Assay

DNA amplification and real-time analysis. The optimized PCR mix (QuantiTect SYBR Green PCR Kit cat. No. 204141) with total volume 20  $\mu$ L is composed of 10  $\mu$ L reaction mix, 0.75  $\mu$ L each of the forward and reverse primers, 4.5  $\mu$ L deionized water, and 4  $\mu$ L DNA. Optimized PCR conditions include 15 min initial denaturation at 95°C, followed by 40 cycles of 5 sec denaturation at 94°C, 10 sec annealing at 55°C and 20 sec extension at 72°C. The minimum DNA template concentration used for amplification was 30 ng/ $\mu$ L. Cut-off Ct value based on the no-template GAPDH control was 32.94. There were no replicates done due to limited sample volumes. Housekeeping primers used were GAPDH forward, 5'-ACCCACTCCTCCACCTTTG-3';

and GAPDH reverse, 5'-CTCTTGCTCTTGCTGGG-3' (Cao, et al., 2008). The real-time PCR reactions were performed using Rotor-Gene Q (2011 model) high precision Real Time thermocycler by Qiagen. Amplification was confirmed using gel electrophoresis. Gel electrophoresis and extraction. Gel was run using 2.5% agarose in 1x TBE buffer, at 130V, 110A for 30 min. Biotium Gel Red Nucleic Acid stain and Amresco ladder EZ-Vision 100 bp were used. The concentration of agarose (2.5%) was selected according to the range of effective separation, since the lengths were around 200bp and differences were between 20-50bp. Gel extraction was done using QIAquick Gel Extraction Kit. DNA sequencing. Samples positive for amplification from the melt-curve analysis that had clear bands were gel extracted and sent to Macrogen, Korea for sequencing. Duplicate samples from the leftover original amplicon were also sent. No concentrating methods were done for the gel extracted samples

### 2.4 Melt-Curve and Sequencing Results Analyses

Melt-Curve Analysis. Batch real-time PCR was performed on 17 samples using 5 primer pairs, A\*02, A\*24, B\*15, C\*07 and DRB\*15. From the high-resolution melt curve analysis, peak temperatures measured up to second decimal place were not enough to distinguish different alleles as confirmed by subsequent sequencing and were much more dependent on the amplicon length. Melt-curve profiles were analyzed using Rotor-Gene Q Series Software 2.1.0. Melt curve peak temperature is a common parameter which may differ for every amplicon and may be considered in analyzing melt curve profiles. (Dwight et al., 2011; Bruzzone et al., 2013)

Sequencing Results Analysis. BLAST was performed using the IMGT/HLA database. Pairwise sequence alignments were performed using EMBOSS-Needle version in 2016 ([https://www.ebi.ac.uk/Tools/psa/emboss\\_needle/nucleotide.html](https://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html)), (Li et al., 2015) and multiple sequence alignments using Clustal Omega version in 2016 (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) (Li et al., 2015).

## 3. RESULTS AND DISCUSSION

### 3.1 Selection of Supertypes

The dbMHC database provides the most data on allele frequencies from a representative Filipino population. The only study on Filipino alleles was from Erlich et al., 2007, which provided a total sample size of 94. Only the top three alleles in terms of frequency for Class I and Class II HLA were chosen for the study due to the limitations in the budget. The alleles and their corresponding supertypes chosen were: A\*02 (A2), A\*11 (A3), A\*24 (A24), B\*15 (B27), C\*07 (C1), DRB1\*15 (DR1) and DQB1\*05 (DQ1). **Table 1** shows the selected Class I and II alleles, their frequencies, and their corresponding supertypes.

**Table 1.** Allele Frequencies for the Filipino Population (Erlich et al., 2007).

Allele	Frequency (N=94)	Supertype
A*02	0.106	A2
A*11	0.266	A3
A*24	0.383	A24
B*15	0.229	B27
C*07	0.362	C1
DRB1*15	0.452	DR1
DQB1*05	0.511	DQ1

### 3.2 Primer Design

After primer assessment and alignment in the IMGT/HLA database, only widely spaced single mismatches between closely related alleles were found. False priming is unlikely even with the presence of 3'-end mismatches (Lefever et al., 2013). Studies have shown that mismatch specificity is only present during the first few cycles (Lefever et al., 2013; Liu, et al. 2012). A single mismatch at the 3'-end has little or no impact on yield, and there must be at least 4 mismatches to completely block the reaction (Lefever et al., 2013; Liu, et al. 2012).

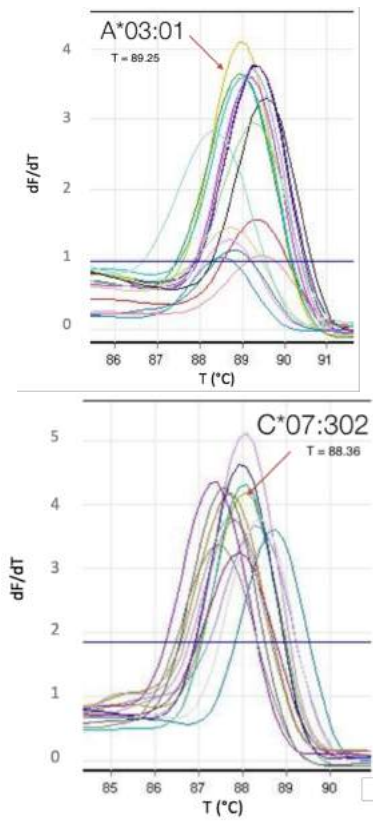
The study used high resolution melting curve analysis to resolve this problem. The specificity in identity matching will depend on the melt-curve profile and not on whether there was an amplification or not.

**Table 2.** Designed Primers and Their Amplicon Lengths

Allele	Primer	Sequence (5' → 3')	Amplicon (bp)
A*02	F	TTCTTCACATCCGTGTCCCG	199
	R	GAGTCTGTGAGTGGGCCTTC	
A*24	F	TTTCTCCACATCCGTGTCCC	196
	R	CTGTGAGTGGGCCTTCACTT	
B*15	F	CCCAGTTCGTGAGGTTCCGAC	178
	R	GCCTCGCTCTGGTTGTAGTA	
C*07	F	GGTCTCACACCTCCAGAGG	169
	R	AACTTGCCTGGGTGATCTG	
DRB1*15	F	TGGCAGCCTAAGAGGAAGTG	187
	R	CCTGCTCCAGGATGCCTTC	
DQB1*05	F	TGTGCTACTTCACCAACGGG	220
	R	TACGCCACCTCGTAGTTGTG	

### 3.3 Real-Time PCR Assay and Melt-Curve Analysis

Amplification was confirmed using gel electrophoresis, which showed clear, distinct, single bands, as shown in Figure 2. As can be seen, the amplicon lengths are consistent with the expected lengths of ~200 bp, no multiple bands are present, and all samples had amplicons.



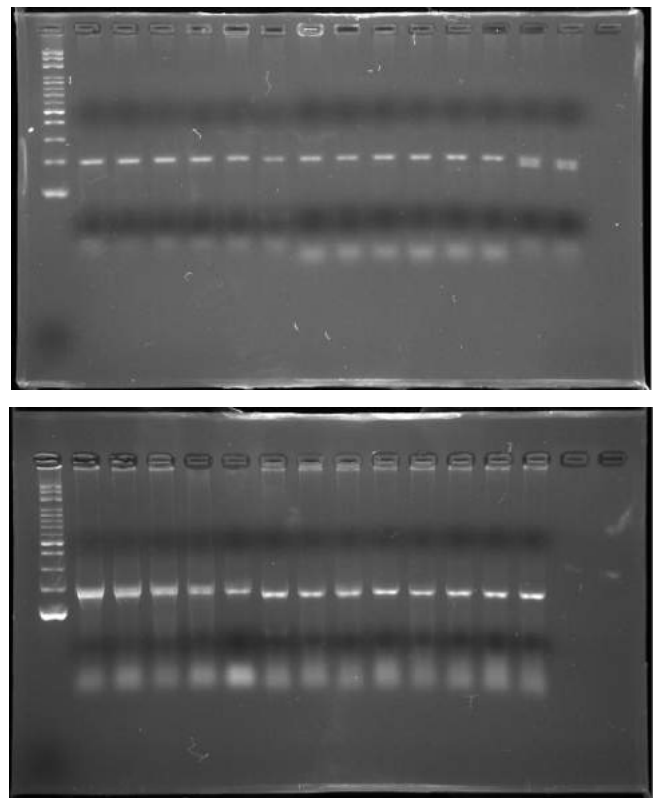
**Figure 1.** Melting Curves of A\*03:01 and C\*07:302 showing identity upon superimposition with other amplicons in a similar allele group. Top: melting curve of A\*03:01 (arrow) superimposed with other amplicons for primer A\*02; Bottom: melting curve of C\*07:302 (arrow) superimposed with other amplicons for primer C\*07. All other melting curves have inconclusive sequencing results (see Appendix B for more details).

Previous studies have shown that melting curves have different shapes for different specific sequences (Zhou, et al. 2004; Reed and Wittwer, 2004). It is theorized that DNA may be assuming an intermediate state during the melting process, as regions of the amplicon that are more stable do not melt immediately (e.g. G-C rich regions) (Huguet et al., 2010).

All melt curves obtained showed unique profiles as seen in **Appendix A**, suggesting that the HLA alleles of the 17 patient samples were all unique. Also, because of mis-priming due to suboptimal PCR conditions, different PCR products tend to have heterozygous sequences, which can have significant effects on the shape of their melt curve. This is consistent with the single bands seen in the agarose gel electrophoresis.

### 3.3 Sequencing Results

The actual HLA alleles of the samples were not previously identified, their PCR products were gel extracted and sent for sequencing. Selected samples



**Figure 2.** Sample agarose gel electrophoresis for sample codes AA (top) and AB (bottom). Leftmost – 100 bp ladder. (top after ladder: AA: 2, 3, 5, 10, 11, 15, 18, 19, 21, 22, 27, 28, AB1,2; bottom after ladder: AB: 3, 10, 13, 16, 17, 21, 22, 24, 26, 27, 28, 31, 32)

with different melting peak temperatures were sent for sequencing. After performing BLAST using the IMGT/HLA database, HLA allele types to the medium-resolution level were not identified for most samples, despite clear non-overlapping chromatograms. There have been inconsistencies with the sequences for some samples (<75% pairwise sequence alignment identity between forward and reverse sequences, and different allele hits), which could reflect poor sample quality. (See **Appendix B**) Many other samples were only identifiable at the low-resolution level, i.e. only up to the HLA allele group (e.g. HLA-A\*02) vs medium resolution which also identifies the specific HLA protein (e.g. HLA-A\*02:101). Although % identity for the top hit results from available sequences in the BLAST database were generally high (>90%), specific HLA proteins either could not be identified or there were multiple hits within the same allele group, rendering the results to be inconclusive. This may be due to some alleles having almost perfectly similar sequences in the amplified region, considering the amplicon lengths were all less than 200 bp, which shows some weakness in the primer design. Heterozygous sequences were also seen in some chromatograms. Although multiple sequence alignment showed conserved regions, the primers were not specific enough to be used for sequence-based typing,

and therefore the samples will need to be typed using another method as a standard. On the other hand, there were two alleles identified at the medium-resolution level using this method, A\*03:01 and C\*07:302. **Figure 1** shows their specific melt curve profile.

With further validation through replication and resequencing using other positive controls, while keeping other factors such as DNA template concentration and purity constant, this method can be used as a tool for possible HLA typing of allogeneic cell therapy specimens.

#### 4. CONCLUSION AND RECOMMENDATIONS

The study explored the use of real-time PCR and melting curve analysis in describing some common HLA Class I alleles in Filipinos. A real-time PCR assay for six selected HLA alleles was optimized. However, only 2 out of 17 samples were identified at the HLA allele and protein level, namely, HLA A\*03:01 and C\*07:302, which were confirmed by sequencing. This may be due to highly similar sequences in the amplified region and heterozygous amplification. More specific primers may be designed to address this problem. Melt-curve peak temperature alone is not enough to differentiate HLA alleles so the actual profiles may need to be compared. The findings in this study will need further validation through replication and sequencing to establish the melting curve profiles for HLA identification use.

#### REFERENCES

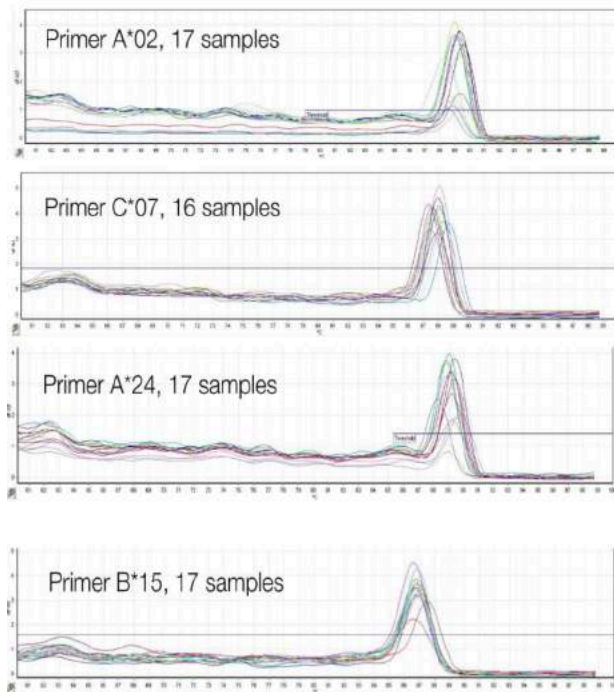
1. Allard M, Oger R, Benlalam H, Florenceau L, Echasserieu K, Bernardeau K, Labarrière N, Lang F, & Gervois N. 2014. Soluble HLA-I/peptide monomers mediate antigen specific CD8 T cell activation through passive peptide exchange with cell-bound HLA-I molecules. *J Immunol.* 192 (11):5090–7. doi:10.4049/jimmunol.1303226
2. Bruzzone CM, Tawadros PS, et al. (2013) Enhanced Primer Selection and Synthetic Amplicon Templates Optimize High-Resolution Melting Analysis of Single-Nucleotide Polymorphisms in a Large Population. *Genet Test Mol Biomarkers*, 17 (9):675–680.
3. Cao Y, Huschtscha L, Nouwens A, Pickett H, Neumann A, Chang A, Toouli C, Bryan T, Reddel R. 2008. Amplification of hTERT in human mammary epithelial cells with limiting hTERT expression levels. *Cancer Res.* 68(9):3115-23.
4. Choi S, Levine J, & Ferrara, J. 2009. Management of Graft-versus-Host Disease. *Immunol Allergy Clin N Am* 30 (2010) 75–101. doi:10.1016/j.jiac.2009.10.001 immunology.theclinics.com
5. Choo SY. 2007. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Medical Journal.* 48(1):11-23.

6. Crux NB and Elahi S. 2017. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? *Front Immunol.* 2017 Jul 18;8:832. doi: 10.3389/fimmu.2017.00832.
7. Doytchinova, I, Guan, P, & Flower, D. 2004. Identifying Human MHC Supertypes Using Bioinformatic Methods. *Journal of Immunology.* 172:4314–4323.
8. Dwight Z, Palais R, Wittwer CT. (2011) uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics.* 27(7):1019–1020.
9. Erlich HA, Alejandrino M, Pozzili P, Panelo A, & Bugawan TL. 2007. Filipino from Luzon Island, Philippines. *Anthropology/human genetic diversity population reports.* In: J.A. Hansen, ed. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference.* I:620–621.
10. Gersuk, V & Nepom, G. 2006. A real-time PCR approach for rapid high resolution subtyping of HLA-DRB1\*04. *J Immunol Methods.* 317(1-2):64–70.
11. Huguet, J. M., Bizarro, C. V., Forns, N., Smith, S. B., Bustamante, C., & Ritort, F. (2010). Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 107(35), 15431–15436. <https://doi.org/10.1073/pnas.1001454107>
12. Leddon SA, Sant AJ. 2010. Generation of MHC class II-peptide ligands for CD4 T-cell allorecognition of MHC class II molecules. *Curr Opin Organ Transplant.* 15(4):505–11. doi:10.1097/MOT.0b013e32833bfc5c
13. Lefever S, Pattyn F, Hellemans J, & Vandesomepele J. 2013. Single-Nucleotide Polymorphisms and Other Mismatches Reduce Performance of Quantitative PCR Assays. *Clinical Chemistry.* 59,10:1470–1480.
14. Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y. M., Buso, N., & Lopez, R. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research*, 43(W1), W580–W584. <https://doi.org/10.1093/nar/gkv279>
15. Liu J, Huang S, Sun M, Liu S, Liu Y, Wang W, Zhang X, Wang H, & Hua W. 2012. An improved allele-specific PCR primer design method for SNP marker analysis and its application. *Plant Methods.* 8:34.
16. Lyubchenko YL, Frank-Kamenetskii MD, Vologodskii AV, Lazurkin YS, & Gause GG Jr. 1976. Fine Structure of DNA Melting Curves. *Biopolymers.* 15:1019-1036.
17. Paunic V, Gragert L, Madbouly A, Freeman J, Maiers M. 2012. Measuring Ambiguity in HLA

Typing Methods. PLoS ONE 7(8): e43585. doi:10.1371/journal.pone.0043585

18. Perng CL, Chang LF, Chien WC, Lee TD, & Chang JB. 2012. Effectiveness and limitations of resolving HLA class I and class II by heterozygous ambiguity resolving primers (HARPs) — a modified technique of sequence-based typing (SBT). *Clinical Biochemistry*. 45:1471–1478.
19. Reed G and Wittwer C. 2004. Sensitivity and Specificity of Single-Nucleotide Polymorphism Scanning by High-Resolution Melting Analysis. *Clinical Chemistry* 50:10: 1748–1754.
20. Smith CJ, Osborn AM. Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol Ecol*. 2009 Jan;67(1):6-20. doi: 10.1111/j.1574-6941.2008.00629.x. PMID: 19120456.
21. Ye et al. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 2012 13:134.
22. Zhou L, Vandersteen J, Wang L, Fuller T, Taylor M, Palais B, & Wittwer CT. 2004. High-resolution DNA melting curve analysis to establish HLA genotypic identity. *Tissue Antigens*. 64:156–164.

#### APPENDIX A. RAW MELTING CURVE PROFILES



#### APPENDIX B: SEQUENCING SUMMARY DATA

Primer	Sample (DNA#)	Length	Melt Peak Temp	% Identity	Top Hit	
A*02	AA1	173	89.25	100	Axx	
	(3)	171		100	A03:01	
	AA2	171	89.6	100	A02x	
	(1)	168		100	A02x	
	AA3	145	88.9	100	A24:29	
	(11)	165		100	Axx	
	AA4	147	89.4	100	A24x	
	(16)	167		100	A03:01	
	AA17	190	89.25	95	A02x	
	(17)	174		94	A02x	
	Y1	150	89.6	100	A24x	
	(2)	169		100	A03x	
	Y2	145	89.65	100	A24x	
	(8)	170		100	Axx	
	A*24	AA10	183	89.1	93	A02x
		(10)	173		97	A02:01
		Y3	144	89.6	100	A24x
(2)		164		97	A68:142	
Y4		144	89.85	99	A24x	
(8)		164		99	A24x	
B*15		AB3	100	87.2	97	B27:109
		(1)	156		93	B37:06
	AB13	68	87.85	96	Cxx	
	(6)	151		98	Gxx	
	AB14	63	87.35	97	Bxx	
	(3)	145		92	Bxx	
	Y5	100	87.1	xxx	xxx	
	(2)	153		95	B37:52	
	Y6	100	87.15	xxx	xxx	
	(8)	163		96	Bxx	
	C*07	AB21	144	88.1	100	Bxx
(11)		100		xxx	xxx	
AB22		145	88.36	100	C07x	
(10)		144		100	C07:302	
AB24		100	89.11	98	Axx	
(7)		141		99	A03:95	
AB31		143	87.85	100	Cxx	
(19)		141		97	C04x	
AB32		100	87.95	96	Bxx	
(18)		100		xxx	xxx	
Y7		143	87.75	100	B40x	
(2)	145		95	C04x		
Y8	141	87.6	100	C08x		
(8)	143		98	C08x		

**Note:**

*Top Hit:* Sequence similarity top hit from BLAST – IMGT/HLA database (top: forward, bottom: reverse)

*x:* specific HLA protein could not be identified

*xx:* HLA group could not be identified

*xxx:* no significant matches found

*% identity:* percent extent to which the sample sequence and the top hit sequence have the same alignment residues (top: forward, bottom: reverse)

## APPENDIX C: SEQUENCES

### AA1

(F)GNNNGNANCGCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTTCGT  
GCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGACCGCGGGCCGCGTG  
GATAGACAGGAGGGCCGAGTATTGGACCGGAACACACGGAATGTGAAG  
GCCACTCACAGACTACCAC  
(R)GTTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
CCCGTGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCC  
GCGAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGGCC  
GGAGTATGACCGAANCCNGN

### AA2

(F)GNNNGNNGGGAACGCTTCTCGCAGTGGGCTACGTGGACGACACGCAGTTC  
GTGCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCG  
TGGATAGAGCAGGAGGTCCGGAGTATTGGACGGGGAGACACGAAAGTG  
AAGGCCACTCACAGACT  
(R)TTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
CAGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCCG  
GATCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGTCCGGA  
GTATGGACGNNANG

### AA3

(F)GNCCGGNACNCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTTCGTG  
GGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCGTGG  
TAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACACGGAAA  
(R)TTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
CAGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCCG  
GATCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGTCCGGA  
GTNTGNCNNNN

### AA4

(F)GGCCGGNCCGCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTTCGT  
GCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCGTG  
GATAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACAGGGAAA  
(R)TTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
CCGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCCG  
CAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGGCC  
GAGTNTGACGANNN

### AA17

(F)NCNNTNNGNCCNGCTTCCNCGCAGTGGGCTACGTGGACGACACGCAGTC  
CCGTGCGGATCGACAGCGACCCGCGAGCCAGAGCATGGAGCCGCGGGCC  
CGTGGATAGAGCAGGAGGGTCCGGAGTATTGGACGGGGAGACACGGAAA  
TGAAGCCCACTCACAGACTACGGGACACGGAAGTGA  
(R)GAGTCTGTGAGTGGGCTTTTACATCCGTGTCCCGCNGCNGCGGGGAG  
CCCCCTTATCGCAGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTGA  
CAGCCGCGGAGCCAGAGGAGGGACCGGGCCCGTGGATAGAGCAGGA  
GGTCCGGAGTNTGACAGANGN

### Y1

(F)CNGNGNCGGGAACGCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTT  
CGTGGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCG  
GTGATAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACAGGGAAA  
(R)TTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
GCCGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCC  
GCGAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGGCC  
GGAGTNTGNCNGANNAGN

### Y2

(F)GGGGNNGNACCCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTC  
GTGCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCG  
TGGATAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACAGGAAA  
(R)TTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
CCGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCC  
GAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGGCC  
GGATATGACGAGANNNGT

### AA10

(F)CNNNNGNACCANNCTCATCGCAGTGGCTACGCGGACGACACGCAGTTCGT  
GCCATCGACAGCGACGCCGCGAGCCAGAGGACGGAGCCGCGGGCCCGTGA  
TAGAGCAGGAGGTCCGAGTATTGGACGGGAGACACGGAAGTGAAG  
CCCACTCACAGAGGACCGGATGTGAGAAA  
(R)TCTGTGAGTGTGCTTACTTCTACACCGTTCCAGCCGGGGCCGGGACC  
CCGTTTATCGCAGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGAC  
ACGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGA  
GGTCCGGAGTNNACNNN

### Y3

(F)GNNNNGNACCCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTTCGT  
GCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCGTG  
GATAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACAGGG  
(R)TTTGTATTTACAATAACTTAAGTCCGTCGAAAAGAGCTTCTTATCGCCGTG  
GGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCCGCGAGC

CCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGCCGGAGTAN  
TGNNNNNNNN

### Y4

(F)GNNNNGNCCCGCTTCTCGCCGTGGGCTACGTGGACGACACGCAGTTTCGT  
GCGGTTTCGACAGCGACGCCGCGAGCCAGAGGATGGAGCCGCGGGCCGCGTG  
GATAGAGCAGGAGGGCCGGAGTATTGGACGAGGAGACAGGG  
(R)TTTTCTTCACATCCGTGTCCCGGCCCGGGCCGCGGGGAGCCCCGTTTATC  
GCCGTGGGCTACGTGGACGACACGCAGTTCTGTGCGTTTCGACAGCGACGCC  
GCGAGCCAGAGGATGGAGCCGCGGGCCCGTGGATAGAGCAGGAGGGCCG  
GAGTANGNNANN

### AB3

(F)GNNTNNTTCGAGGATGGAGCGGGCCCGTGGATAGAGCAGGGAGGGG  
CCGGAGTATTGGACCGAGGAGACCGATCTGCAAGGACCCCGCACAGAC  
(R)TCCAGTTTGTAGGTTTTCGACAGCGACTCGGCGAGTCCGAGGATGGAGCC  
GCGGGCCCGTGGATGGAGCAGGAGGGCCGGAGTATTGGACCGGGGAG  
ACACAGATCGTCAAGGCCACGGCACAGACTGACACGANNAACNNTNGAT  
NTNN

### AB13

(F)NNNNGGTCAGGGAGNAGCGGGCCCGTGGANGGGCAGGAGGGGCC  
GGAGTATTGGACCGGGAGA  
(R)TCCAGTTTCGTGAGTTTCGACAGCGACTCGGCGTGTCCGAGGATGGAGCC  
GCGGGCCCGTGGTGGAGCAGGAGGGCCGGAGTATTGGACCGGGGAG  
ACACGAGATCGTCAAGGCCACGCACAGACTACAGAGAGANCCGNNNNN

### AB14

(F)GNNNNNNNCGAGGAGGAGCGCGGGCCCGTGGATAGAGCAGGAGGGCC  
CGGAGTATTGCGACC  
(R)TCCAGTTTCGTGAGTTTCGACAGCGACCGCGAGTCCGAGGATGGAGCC  
GCGGGCCCGTGGTGGAGCAGGAGGGCCGGAGTATTGGACCGGGGAGACA  
CGGACCTGCAAGGCCACGCACAGACTGACCGAAGAACCGGTTNN

### Y5

(F)NNNNNNNNGTCAAAGAGNGCGGGCCCGTGGATAGAACCAGGGAGGGG  
CCCGGAGTACTGGGACCGGACAGAAGTGTGTTAGTTTTATTTCCGATCG  
(R)TTCCAGTTTCGTGAGTTTTCGACAGCGACTCCGCGAGTCCGAGGATGGAGC  
CGCGGGCCCGTGGTGGAGCAGGAGGGCCGGAGTATTGGACCGGGGAGACA  
GACACAGATCGTCAAGGCCACGCACAGACTACAGGAAGGAACCGCANNNC  
N

### Y6

(F)TGACNNGTCCGAGAGGAGCGGGCCCGTGGATAGANACAGAGGGCC  
GAGAGTATTGGACACGAGGAGACACGGAACGCCAAGACAACTGA  
(R)TCCAGTTTCGTGAGTTTCGACAGCGACCGCGAGTCCGAGGATGGAGCC  
GCGGGCCCGTGGTGGAGCAGGAGGGCCAGAGTATTGGGACCGGGG  
AGGACACGAGGATCATCTGAAGGCCACNACCAGGACTGTCNNAGATGNCC  
CGAANAAGNNA

### AB21

(F)NNGGNCNNGNCCGGACGGGCGCTCCTCCGCGGCATAACCAGTA  
CGCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGACCTGCGCTCTG  
ACCGCCCGGACACCGCGGCTCAGATCACCCAGCGCAAGTACC  
(R)TGTGACGCGTCTCGTCCACATCCGATCGGCTGCGGCGGGGGGGCG  
GCCGGGTTCTGGCCGAACGAGGACCGGCTCCGANCCNNTTNNNNGA

### AB22

(F)GNNANGNNGCACTGGGGGACGGGCGCTCCTCCGCGGATGACCAGT  
CGCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGACCTGCGCTCTG  
GACCGCCGCGGACACCGGCTCAGATCACCCAGCGCAAGTTAT  
(R)GTGGTCTCACACCTTCCAGAGGATGTCTGGCTGCGACCTGGGGCCGAGC  
GGCGCTCCTCCGCGGATGACCAGTCCGCTACGACGGCAAGGATTACAT  
CGCCCTGAACGAGGACCTGCGCTCCTACCGCCGNGATNCCC

### AB24

(F)GNNNTACTGGGNGCGGACGGGCGCTCCTCCGCGGATACCAGCAGGAC  
GCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGACCTGCGCTCTG  
(R)TTGGTCTCACACCTTCCAGAGGATGTATGGCTGCGAGTGGGGTCCGAGC  
GGCGTCTCCTCCGCGGATACCGGACGACGCTACGACGGCAAGGATTACAT  
CGCCCTGAACGAGGACCTGCGCTCCTGNCNNGNNTAGN

### AB31

(F)NNNNNNCACTGGGGCGGACGGGCGCTCCTCCGCGGATGACCAGTCC  
GCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGATCTGCGCTCTGGA  
CCGCGCGGACACCGCGGCTCAGATCACCCAGCGCAAGTTAA  
(R)TTCTCACACATTCAGAGGATGTTTGGCTGCGACCTGGGGCCGAGGGG  
CCCTCCTCCGCGGATGACCAGTTCGCTACGACGGCAAGGATTACATCGC  
CCTGAACGAGGNTCTGCGCTCCTGACCGCCGNNNNNNN

### AB32

(F)CNNGGCCACCGNNGGAGGGCGCTCCTCCGCGGCATAACCAGGTCC  
CCTACGACGGCAAGGATACATCGCCCTGAACGAGGACCTGAGCTCCGGCAC  
(R)AATCGTACCACAAAACCTCCCGCGACCGTAAGCAGCCGCTGACGG  
CAAGGATCCTCGCCCTCACGAGGACCTGGGCTCCTGNAGCCCNCCGT  
Y7



(F)CGNGNCTACTGGGGCGGACGGGCGCCTCCTCCGCGGGCATAACCAAGTTC  
GCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGACCTGCGCTCCTGGA  
CCGCCGCGGACACGGCGGCTCAGATCACCCAGCGCAAGTTAA  
(R)TTGGTCTCACACCTTCTAGAGGATGTATTGGCTGCGACGTGGGGCCGGAC  
GGGCGCCTCCTCCGCGGGCATAACCAAGTTCGCCTACGACGGCAAGGATTACA  
TCGCCCTGANCGAGGATCTGCGCTCCTNGACCGGGNGNNNC

**Y8**

(F)GNNGTNACTGGGGCGGACGGGCGCCTCCTCCGCGGGTATAACCAAGTTCGCC  
TACGACGGCAAGGATTACATCGCCCTGAATGAGGACCTGCGCTCCTGGACCG  
CCGCGGACACGGCGGCTCAGATCACCCAGCGCAAGTAA  
(R)TGGTCTCACACCGTCCAGAGGATGTATGGCTGCGACCTGGGGCCCGACGG  
GCGCCTCCTCCGCGGGCATAACCAAGTTCGCCTACGACGGCAAGGATTACATC  
GCCCTGAATGAGGACCTGCGCTCCTGACCGCGGATCCANN

# ***In Silico* Pathway Analysis of the Anti-cancer Mechanism of Selected Active Components of Virgin Coconut Oil and their Key Targets**

**Excellces Dee Montemayor<sup>1</sup>, Mayrell Ann F. Ravina<sup>1</sup>, Jay T. Dalet<sup>1</sup>, Francisco M. Heralde III<sup>2</sup>**

<sup>1</sup>Department of Biology, College of Arts and Sciences, University of the Philippines Manila, Padre Faura St., Ermita, Manila

<sup>2</sup>Department of Biochemistry and Molecular Biology, College of Medicine, University of the Philippines Manila, Pedro Gil St., Ermita, Manila

## **Email address:**

edmontemayor@up.edu.ph, mfravina@up.edu.ph, jtdalet@up.edu.ph

## **To cite:**

Montemayor ED, Ravina MAF, Dalet JT, and Heralde FIIM.2020. *In Silico* Pathway Analysis of the Anti-cancer Mechanism of Selected Active Components of Virgin Coconut Oil and their Key Targets. PJBMB. Vol. I, No. 1, 2020, pp. 37-53. doi: 10.5555/pjbmb.ph.2020.01.01.37

**Received:** 07 31, 2020; **Accepted:** 12 15, 2020; **Published:** 01 04, 2021

**Abstract:** The Philippines has 40 virgin coconut oil (VCO) producers that accommodate the increasing demand for VCO (Manohar et al., 2007). VCO is an edible lipid-based product extracted from coconut copra by a wet process (Marina et al., 2009) and it is used for producing massage oils, lotions, balms, creams and soaps (Manohar et al., 2007). It is also suggested to be a possible complementary and alternative medicine (CAM) for cancer. This *in silico* study contributes in analyzing the anti-cancer relationship between VCO and its key cancer protein targets, through an integrated approach of bioinformatics and pharmacology. VCO active components were screened by oral bioavailability (OB), drug-likeness (DL), and probable anti-cancer activity while their key targets in humans were determined using z'-scores, cancer pathways involvement, and centrality algorithms. These compounds were docked to determine the active components' putative efficacy as ligands while the top ten KEGG pathways enriched among cancer protein targets were obtained. Primers and probes of key targets for gene expression analysis were also designed. The candidate active components were 2-heptanone, 2-pentanone, and the suggested transcription factor inhibitors are dihydrokaempferol, ferulic acid, and quercetin. They were effective ligands to the key targets: AKT1, HRAS, HSP90AA1, MAPK1, EGFR, MAPK8, RHOA, ESR1, PIK3R1, and MMP9. Moreover, the top enriched pathways were pathways in cancer, focal adhesion, hepatocellular carcinoma, fluid shear stress and atherosclerosis, endocrine resistance, prostate cancer, colorectal cancer, PI3K-Akt signaling pathway, proteoglycans in cancer, and Ras signaling pathway. The interactions of VCO and key targets in these pathways suggested mechanisms that may be highly essential for treating hepatocellular, prostate and colorectal cancers. Ultimately, this study provides leads for focusing future *in vitro* studies on the anti-cancer activity of VCO components in cells.

**Keywords:** VCO, active components, key targets, cancer, AKT1, quercetin, PI3K-Akt signaling pathway, docking

## **1. INTRODUCTION**

Cancer is the second leading cause of death internationally (World Health Organization, 2018) which caused 10.4% of the total deaths in the Philippines last 2016 (Bersales, 2018). Because of exorbitant leading cancer treatments, complementary and alternative medicine (CAM) has gradually become an alternative

choice for managing cancer patients (Dai et al., 2018). Virgin coconut oil (VCO) was shown to have anti-cancer activity against HepG2 liver and KB oral cancer cells (Verma et al., 2019). It is an edible lipid-based product derived from coconut copra. Unlike refined coconut oil, virgin coconut oil is extracted by a wet process that retains most of its therapeutic components (Marina et al., 2009). Because VCO is used in producing food

supplements, massage oils, lotions, balms, creams and soaps, it has an increasing demand. It was a Philippine phenomenon of the 21st century with over 40 producers in the Philippines that are mostly based in the National Capital Region and Southern Luzon (Manohar et al., 2007). Although some components of VCO have already been determined such as its monoglyceride, diglyceride, fatty acid, volatile organic compound, phenolic, sterol, and tocopherol content, the active components of VCO responsible for its anti-cancer activity remain unclear (Verma et al., 2019). Bioinformatics and pharmacology are *in silico* approaches (Dai et al., 2018) essential in the development of drug discovery (Chordia & Kumar, 2018). However, *in silico* studies on VCO against cancer are lacking. Hence, this study was undertaken to analyze *in silico* the possible anti-cancer mechanism of selected VCO active components to their key cancer protein targets. Specifically, the study aims to identify candidate active compounds from VCO through virtual screening in terms of oral bioavailability, drug-likeness, and probable anti-cancer activity; to determine *in silico* the key cancer protein targets of the selected active components of VCO using z'-scores, the KEGG human pathways in cancer, and centrality measures; to determine an optimized dock conformer to the key cancer protein targets among the selected active VCO components; to determine *in silico* the biological pathways linked to the cancer protein targets of selected active components of VCO; and to design *in silico* gene expression oligonucleotide primers and probes for the targeted key cancer protein-coding genes.

## 2. METHODOLOGY

### 2.1. Virtual Screening for VCO Active Components

#### 2.1.1. Creating database of VCO chemical components

Data on VCO composition was collected from Google Scholar and NCBI PubMed databases using the general keyword "virgin coconut oil." The compounds were classified chemically and the canonical Simplified Molecular-Input Line-Entry System (SMILES), 2D structures, and 3D structural data format (SDF) of these compounds and the selected controls: Chlorambucil, 1,2-dioleoyl-rac-glycerol, 3-Amino-N-(4-fluorophenyl)-6-(2-thienyl),-4-(trifluoromethyl)thieno[2,3-b] pyridine-2-carboxamide (FDI-6), and Ellipticine were obtained from NCBI PubChem database. Chlorambucil is an oral anti-cancer drug (Wishart et al., 2006) while 1,2-dioleoyl-rac-glycerol is a diacylglycerol not used on humans (Cayman, n.d.). Because of these characteristics, these two compounds were suitable positive and negative controls for oral bioavailability, drug-likeness, and anti-cancer activity, respectively. FDI-6 is a transcription inhibitor while Ellipticine is a transcription stimulant (Lambert et al., 2018). Thus, they were selected as suitable positive and negative controls for transcription inhibition, respectively.

#### 2.1.2. Screening for VCO active components by OB

After entering the canonical SMILES of each VCO component into SWISSADME (<http://www.swissadme.ch>), compounds that obtained an oral bioavailability (OB) score >0.50 and satisfied the Lipinski's, Veber's, and Egan's (LVE) rules without violations were selected for DL-screening. This threshold was selected based on the use of F =0.50 as a binary classification threshold (Kim et al., 2014). Chlorambucil and 1,2-dioleoyl-rac-glycerol served as the positive and negative control, respectively.

#### 2.1.3. Screening for VCO active components by DL

After uploading the SDF files of each OB-screened VCO component into Drug-likeness Tool (DruLiTo) (Sharma et al., 2014), compounds with a weighted and unweighted quantitative estimate of drug-likeness (QED) >0.35 were selected. This threshold was selected based on the value at which the mean QED of unattractive complex compounds is lower; QED surpasses Lipinski's rule; and Veber comes close to QED (Bickerton et al., 2012). The positive and negative controls used were chlorambucil and 1,2-dioleoyl-rac-glycerol, respectively.

#### 2.1.4. Screening for VCO active components by Pa of anti-cancer activity

After entering the canonical SMILES of each OB- and DL-screened VCO component into Prediction of Activity Spectra for Substances or PASS (<http://www.pharmaexpert.ru/passonline/>), compounds with a probability of activity (Pa) value >0.6 as the threshold set by Hashemi et al. (2014) for anti-cancer (anticarcinogenic, and antineoplastic) activity were selected as active components. Selected compounds with a Pa >0.4 as the chosen threshold for transcription factor (TF) inhibition were noted. This threshold was selected based on the fact that greater Pa values have lesser probability of having false positive results (Basanagouda et al., 2011), and higher Pa values have a higher percentage of missed active compounds (Poroikov, Filimonov, & Associates, n.d.). The positive and negative controls were chlorambucil and 1,2-dioleoyl-rac-glycerol for anti-cancer activity and FDI-6 and ellipticine for TF inhibition, respectively.

## 2.2. Determining Key Cancer Protein Targets

### 2.2.1. Identifying human protein targets

The SDF files of each selected VCO active component were then uploaded into PharmMapper (<http://www.lilab-ecust.cn/pharmmapper/>). This software predicted up to 300 human protein targets for each active component of VCO. These protein targets were generated with corresponding z'-scores (PharmMapper, 2019). The human protein targets with negative z'-scores were disregarded while the remaining targets with positive z'-scores were considered significant protein targets of the screened VCO active components. This threshold was based on the certainty that large positive z'-score indicates high significance of the target to the query compound and that large negative z'-score indicates that the target is

not significant (PharmMapper, 2019). The HUGO Gene Nomenclature Committee (HGNC) symbols of these targets were obtained from Ensembl (<http://useast.ensembl.org/>).

### **2.2.2. Screening for cancer protein targets by involvement to cancer**

After entering the HGNC symbols of the selected human protein targets into STRING (<https://string-db.org/>) for protein to protein interaction (PPI) network mapping, targets analyzed to have involvement in the KEGG human pathways in cancer were selected as cancer protein targets and kept as input nodes of the network. Targets without involvement were removed and the tab separated values (TSV) file of the cancerous PPI network was downloaded. The TFs among the cancer protein targets were determined through Ensembl BioMart using the Gene Ontology (GO) term accession: GO:0003700 filter for TFs (Romero et al., 2005).

### **2.2.3. Screening for key targets by centrality**

After uploading the TSV file of the cancerous PPI network into Cytoscape v3.7.2. (Shannon et al., 2003), the top ten cancer protein targets with the highest degree centrality, closeness centrality, and betweenness centrality upon centrality analysis of the network were selected as key cancer protein targets. The SDF files of these key targets were obtained from NCBI Pubchem database.

## **2.3. Generating Optimized Dock Conformers of VCO Active Components and Key Targets**

### **2.3.1 Docking VCO active components and key targets**

After the SDF files of the VCO active components (ligands) and their respective key targets (receptor proteins) were converted to PDBQT formats using OpenBabelGui (O'Boyle et al., 2011), they were uploaded into AutoDock (Morris et al., 2009) for docking. The conformation with the highest ligand-binding affinity (kcal/mol) or most negative change in Gibbs free energy ( $\Delta G$ ) generated by AutoDock Vina for each protein-ligand match were selected. However, all conformations with  $-\Delta G$  values were considered to have high binding affinity while  $+\Delta G$  values were considered low.

## **2.4. Determining Biological Pathways Linked to Cancer Protein Targets**

### **2.4.1. Constructing VCO compound-cancer target network**

After uploading the TSV file of the cancerous PPI network into Cytoscape v3.7.2, the nodes and respective edges of VCO active components were mapped into the existing network to obtain a VCO compound-cancer target network (Li et al., 2018).

### **2.4.2. Pathway enrichment analysis of cancer protein targets**

After entering the HGNC symbols of the cancer protein targets into WEB-based GENE SeT AnaLysis Toolkit or

WebGestalt over-representation analysis (ORA) (Liao et al., 2019) for KEGG human pathway enrichment analysis, the top ten pathways with the lowest adjusted P value  $<0.01$  were selected (Dai et al., 2018). This threshold is based on the fact that a smaller P value dictates stronger significance (McLeod, 2019).

## **2.5. Designing Primers and Probes of Targeted Key Cancer Protein-Coding Genes**

### **2.5.1. Primer design of key targets**

After the shortest exon-exon junction sequence template was obtained from the longest protein coding gene of each key target through Ensembl (Yates et al., 2020), the corresponding primers with a product size of 70-200 base pairs (bp) (Sadangi, 2015), primer size of 18-30 bp (Abd-Elsalam, 2003), T<sub>m</sub> of 57-63°C (Sadangi, 2015), max T<sub>m</sub> difference of 2 (Prediger, 2013), GC content of 40-60% (Alfandary, 2015), max hairpin melting temperature of 50°C (Abd-Elsalam, 2003), max poly-x of 1, and max self-complementarity of 0 (Sadangi, 2015) were obtained using Primer3Plus (<https://primer3plus.com/>). When no primers were generated, the settings or template were adjusted.

### **2.5.2. Selecting primers**

The primers were screened for uniqueness to the desired sequence (Prediger, 2013) in humans using the BLAST option of Primer3Plus which links to NCBI nucleotide BLAST. Primers with an exon-exon junction overlap (Vandenbroucke et al., 2001), GC clamp (Abd-Elsalam, 2003), a query coverage of 100% (Sadangi, 2015) and E-value  $<1$  were selected. This threshold was set to limit the number of hits seen by chance, ensure that the match is significant (Altschul et al., 1997), and avoid off-target interactions (Prediger, 2013).

### **2.5.3. Probe design of key targets**

After entering the designed primer sequence for each key target into PrimerQuest Tool (<https://www.idtdna.com/primerquest/>), the probes with an amplicon size of 70-150 bp, probe length of 20-30 bp, T<sub>m</sub> 6-8°C higher than the primers, and GC content of 35-65% were obtained. When no probes were generated, the settings were adjusted. The probe annealing temperatures and  $\Delta G$  values of self-dimers, hairpins, and heterodimers were obtained from NEB T<sub>m</sub> calculator (<https://tmcalculator.neb.com>) and UNAFold program (<https://www.idtdna.com/unafold/>), respectively.

### **2.5.4. Selecting probes**






The probes were screened for uniqueness to the desired sequence in humans using NCBI nucleotide BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). Probes with close proximity to the primer without overlapping the primer-binding site, an annealing temperature 5°C below the primer T<sub>m</sub>,  $\Delta G$  of self-dimers, hairpins, and heterodimers  $>-9.0$  kcal/mole (Prediger, 2013), a query coverage of 100% (Sadangi, 2015), and E-value  $<1$  (Altschul et al., 1997) were selected.

### 3. RESULTS AND DISCUSSION

#### 3.1 VCO Active Components

Out of 40 VCO chemical components, 23 passed OB-screening, 21 passed subsequent DL-screening, and only 5 out of 21 had probable anticancer activity. Volatile organic compounds: 2-heptanone, and 2-pentanone (Santos et al., 2011), and phenolic compounds: dihydrokaempferol, ferulic acid, and quercetin (Illam et al., 2017) with OB >0.50, DL >0.35 without LVE rule violations, and Pa >0.60 for antineoplastic/anticarcinogenic activity (Table 1) were selected as active components.

Table 1. Screened VCO Active Components. Table shows the VCO active components, 2D structures, their oral bioavailability (OB) score, drug-likeness (DL) via weighted QED (W), and probable (Pa) anticancer (anticarcinogenic and antineoplastic), and TF inhibitory activities.

#	Active Component	Chemical Structure	OB	DL	Anti-carcinogenic Activity	Anti-neoplastic Activity	TF Inhibition
					Pa	Pa	Pa
Volatile Organic Compounds							
1	2-Heptanone		0.55	0.507	0.261	0.698	0.332
2	2-Pentanone		0.55	0.496	0.259	0.715	0.347
Phenolic Compounds							
3	Dihydrokaempferol		0.55	0.639	0.792	0.715	0.417
4	Ferulic Acid		0.55	0.748	0.616	0.601	0.417
5	Quercetin		0.55	0.458	0.757	0.797	0.583

OB signified the ability of a compound to systematically circulate in the body after being orally taken (Tian et al., 2011) which was indicated by OB scores (Martin, 2005) and flagged by LVE rules (Lagorce et al., 2008). LVE rules state that compounds should have  $\leq 5$  hydrogen bond donors,  $\leq 10$  hydrogen acceptors,  $< 500$  Da molecular mass, and  $\leq 5$  partition coefficient (Lipinski et al., 2001);  $< 10$  rotatable bonds (ROTB) and  $< 140$  Å polar surface area (PSA) (Veber et al., 2002); and  $0 \geq$  topological polar surface area (tPSA)  $\leq 132$  Å<sup>2</sup> and  $-1 \geq \log P \leq 6$  (Craciun et al., 2015), respectively. DL signified the compound's chemical suitability for drug use (Dai et al., 2018) as indicated by weighted/unweighted QED.

QED is the measure of a compound's drug-likeness based on its desirability. It allows compounds to be ranked by the elusive quality of chemical attractiveness (Bickerton et al., 2012). These two parameters distinguished which components give VCO the properties of an oral drug but these parameters fail to distinguish which components have probable anti-cancer activity. Pa values determined the probability to exhibit such activity (Filimonov & Poroikov, 2008). Thus, the VCO components that satisfied all three parameters were considered the active components. Additionally, dihydrokaempferol, ferulic acid, and quercetin with Pa >0.40 for TF inhibition (Table 1) were suggested to disrupt the synthesis or activity of TFs (Poroikov, Filimonov, & Associates, n.d.).

#### 3.2 Screened Key Targets

The predicted protein targets of each VCO active component (approximately 1200 in total) were first evaluated through the value of their generated z'-scores. Large negative z'-scores show low significance of the target to the query compound (PharmMapper, 2019), hence all the targets with negative z'-scores values were not considered significant and excluded in the succeeding screening. The remaining 291 significant protein targets were scrutinized for their involvement in the human pathways of cancer using the KEGG Pathway database. Out of 291 targets, no targets were predicted from 2-pentanone and 49 were cancer protein targets (Figure 1).

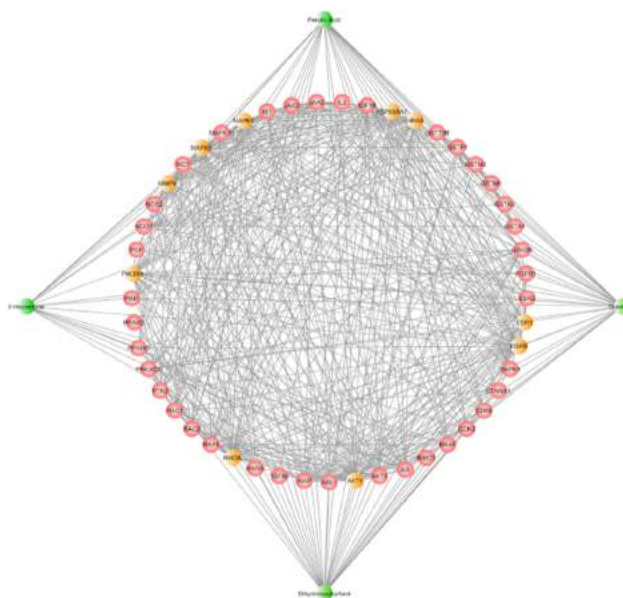


Figure 1. VCO Compound-Cancer Target Network. Image shows the compound-cancer target network of the VCO active components and their corresponding cancer targets. The VCO active components, the cancer targets, and the key targets are represented by green nodes, red nodes, and orange nodes, respectively.

Among them, AR, ESR1, ESR2, PPARD, PPARG, and RXRA were TFs which are effective in the genetic

transcription initiation, stimulation, or termination (Poroikov, Filimonov, & Associates, n.d.) of the target proteins in the pathways of cancer while AKT1, HRAS, HSP90AA1, MAPK1, EGFR, MAPK8, RHOA, ESR1, PIK3R1, and MMP9 with the top 10 highest centrality measures (Table 2) were selected as key targets.

Table 2. Screened Key Targets. Table shows the ten key cancer protein targets and their HGNC symbols, and degree, closeness, and betweenness centrality measures.

#	Key Target	Symbol	Degree Centrality	Closeness Centrality	Betweenness Centrality
1	RAC-alpha serine/threonine-protein kinase	AKT1	34	0.77419355	7.697906E-02
2	GTPase HRas	HRAS	34	0.77419355	5.671742E-02
3	Heat shock protein HSP 90-alpha	HSP90AA1	32	0.75000000	7.845045E-02
4	Mitogen-activated protein kinase 1	MAPK1	32	0.75000000	5.223285E-02
5	Epidermal growth factor receptor	EGFR	31	0.73846154	5.836740E-02
6	Mitogen-activated protein kinase 8	MAPK8	30	0.72727273	5.507305E-02
7	Transforming protein RhoA	RHOA	28	0.70588235	3.972225E-02
8	Estrogen receptor	ESR1	28	0.70588235	3.707980E-02
9	Phosphatidylinositol 3-kinase regulatory subunit alpha	PIK3R1	28	0.66666667	2.873797E-02
10	Matrix metalloproteinase-9	MMP9	25	0.67605634	2.965179E-02





























Centrality algorithms are used to interpret the roles and impact of specific nodes on the entire cancerous PPI network. According to Disney (2020) from Cambridge Intelligence, degree centrality assigns an importance score based on the number of links with the other nodes. Higher values for degree centrality mean that the node is more central (Golbeck, 2015). Betweenness centrality measures the number of times a node lies on the shortest path between other nodes. This measure is used to measure one node which has the mediation role in a network (Zhang & Luo, 2017). It tells us which nodes bridges between nodes in a network. The shortest paths are identified and the number of times each node falls on one is counted (Disney, 2020). The target node which has many short paths would have a high betweenness centrality (Perez & Germon, 2016). Lastly, closeness centrality scores each node based on their closeness to all other nodes in the network. This measure calculates the shortest paths between all the nodes and then assigns each node a score based on the sum of shortest paths (Disney, 2020). If the length of node X's shortest paths with other nodes in the network is small, then this node has a high closeness centrality. (Zhang & Luo, 2017). Thus, the key targets are the proteins with the most crucial role in the cancer protein target interactions.

### 3.3 Docked VCO Active Components and Key Targets

The active components 2-heptanone, dihydrokaempferol, ferulic acid, and quercetin with  $-\Delta G$  scores were considered effective ligands to their respective key targets. Among the ligands, 2-heptanone had the lowest  $\Delta G$  score for MAPK1. Dihydrokaempferol had the lowest  $\Delta G$  score for AKT1, ESR1, and MMP9 and quercetin had the lowest  $\Delta G$  score for HRAS, HSP0AA1, EGFR, MAPK8, RHOA, and PIK3R1. However, 2-heptanone,

dihydrokaempferol, ferulic acid, and quercetin had the lowest  $\Delta G$  scores for MAPK8, MMP9, HSP90AA1, and MAPK8 among their key targets, respectively (Table 3).

Table 3. Docked VCO Active Components and Key Targets in First Conformation. Table shows the binding affinities in kcal/mol and protein-ligand complexes (P-L) of the docked VCO active components (green ligands L) with corresponding key targets (red receptor proteins P) in their first conformation. The "\*" represents cells that were intended to be empty.

#	L	First Conformation							
		P	AKT1	HRAS	HSP90AA1	EGFR	MAPK8	RHOA	ESR1
1	Ferulic Acid	$\Delta G$	-6.6	-5.3	-6.8	-5.8	-6.7	-5.9	-6.3
		P-L							
2	Quercetin	$\Delta G$	-7.5	-8.0	-7.3	-8.4	-6.9	-7.3	-6.6
		P-L							
3	Dihydrokaempferol	$\Delta G$	-8.0	-6.5	-6.8	-8.1	-7.5	-8.3	*
		P-L							
4	2-Heptanone	$\Delta G$	-4.1	-3.7	-4.4	*	*	*	*
		P-L							

The binding of a receptor protein and a ligand occurs only when the value of their system's change in Gibbs free energy ( $\Delta G$ ) is negative (Afriza et al., 2018) which is why conformations with  $+\Delta G$  scores were considered to have low binding affinity while  $-\Delta G$  scores were considered high. Analogous to spontaneous processes, this binding action occurs when the system reaches an equilibrium state at constant pressure and temperature. The  $\Delta G$  reflects the extent of a protein-ligand interaction and the binding affinity of a ligand to a given protein receptor (Bronowska, 2011). This score also shows how stable the resulting complexes of the receptors and ligands are. It is an essential characteristic of effective drugs which occurs when the value of their system's  $\Delta G$  is negative. Thus, more negative  $\Delta G$  values indicated greater binding affinity and complex stability (Afriza et al., 2018).

### 3.4 Top Ten Pathways of VCO Compound-Cancer Target Network

The interaction of the VCO active components and cancer protein targets as visualized in Figure 1 were suggested to impact specific pathways in humans. These pathways (Table 4) are significant in explaining the action of cancer protein targeting by VCO active components. The construction of compound-cancer target networks integrates both the compound to target

protein interactions and the target protein to target protein interactions. The construction of such networks aids in the identification of the mechanism of a potential drug's treatment (Li et al., 2018).

Table 4. Top Ten Pathways of Cancer Protein Targets. Shown in the second column are the names of pathways, followed by the involved cancer protein targets in the third column where the key targets are in bold. C is the gene set size of the number of reference targets in the category; E is the expected number in the category; O is the number of targets in both the gene set and the category that overlap; R is the enrichment ratio; P is the raw p-value upon hypergeometric test for evaluating significance; and F is the false discovery rate or the adjusted p-value by the multiple test adjustment.

#	Pathways	Targets	C	E	O	R	P	F
1	Pathways in cancer	All 49 Cancer Protein Targets	524	3.5313	49	13.876	0	0
2	Focal adhesion	<b>AKT1, AKT2, BRAF, EGFR, GSK3B, HRAS, IGF1R, MAPK1, MAPK10, MAPK8, MET, PGF, PIK3R1, PTK2, RAC1, RAC2, RAF1, RHOA, and XIAP</b>	199	1.3411	19	14.168	0	0
3	Hepatocellular carcinoma	<b>AKT1, AKT2, BRAF, CDK6, EGFR, GSK3B, GSTA1, GSTA3, GSTM1, GSTM2, GSTP1, GSTT2B, HRAS, IGF1R, MAPK1, MET, NQO1, PIK3R1, RAF1, and TGFβ2</b>	167	1.1254	20	17.771	0	0
4	Fluid shear stress and atherosclerosis	<b>AKT1, AKT2, GSTA1, GSTA3, GSTM1, GSTM2, GSTP1, GSTT2B, HSP90AA1, MAPK10, MAPK8, MMP9, NQO1, PIK3R1, PTK2, RAC1, RAC2, and RHOA</b>	138	0.93	18	19.355	0	0
5	Endocrine resistance	<b>AKT1, AKT2, BRAF, EGFR, ESR1, ESR2, HRAS, IGF1R, MAPK1, MAPK10, MAPK8, MMP9, PIK3R1, PRKACA, PTK2, and RAF1</b>	98	0.68043	16	24.227	0	0
6	Prostate cancer	<b>AKT1, AKT2, AR, BRAF, CDK2, EGFR, FGFR1, GSK3B, GSTP1, HRAS, HSP90AA1, IGF1R, MAPK1, MMP9, PIK3R1, and RAF1</b>	97	0.65369	16	24.476	0	0
7	Colorectal cancer	<b>AKT1, AKT2, BRAF, EGFR, GSK3B, HRAS, MAPK1, MAPK10, MAPK8, PIK3R1, RAC1, RAC2, RAF1, RHOA, and TGFβ2</b>	86	0.57956	15	25.882	0	0
8	PI3K-Akt signaling pathway	<b>AKT1, AKT2, CDK2, CDK6, EGFR, FGFR1, GSK3B, HRAS, HSP90AA1, IGF1R, IL2, JAK2, JAK3, KIT, MAPK1, MET, PGF, PIK3R1, PTK2, RAC1, RAF1, and RXRA</b>	354	2.3856	22	9.2218	1.11E-16	4.52E-15
9	Ras signaling pathway	<b>ABL1, AKT1, AKT2, EGFR, FGFR1, HRAS, IGF1R, KIT, MAPK1, MAPK10, MAPK8, MET, PGF, PIK3R1, PRKACA, RAC1, RAC2, RAF1, and RHOA</b>	232	1.5635	19	12.152	2.22E-16	7.24E-15
10	Proteoglycans in cancer	<b>AKT1, AKT2, BRAF, EGFR, ESR1, FGFR1, HRAS, IGF1R, MAPK1, MET, MMP9, PIK3R1, PRKACA, PTK2, RAC1, RAF1, RHOA, and TGFβ2</b>	198	1.3343	18	13.49	2.22E-16	7.24E-15

Pathways in cancer is an overview of the different signaling networks involved in cancer for humans. Because the targets were screened for their involvement in cancer through this pathway, it was evident for all 49 cancer protein targets to be linked to it.

Focal adhesions are specialized structures in the contact points of the cell's extracellular matrix (Kanehisa & Goto, 2000). Targeting focal adhesion proteins in cancer strongly sensitize cancer cells to different cancer treatments such as radiotherapy, chemotherapy, and novel molecular therapeutics (Eke & Cordes, 2015). In focal adhesions, targeting key targets AKT1 and

PIK3R1 of the PI3K-Akt signaling pathway may hinder cell survival. Similarly, targeting HRAS, MAPK1, and MAPK8 involved in the MAPK signaling pathway may constrain cell survival or cell proliferation. EGFR is involved in both of the previously mentioned pathways as a receptor while RHOA is indirectly linked to the regulation of the actin cytoskeleton of the cell. Thus, targeting RHOA may decrease cell motility (Kanehisa & Goto, 2000).

Three pathways were specific types of cancer namely hepatocellular carcinoma, prostate cancer, and colorectal cancer. Hepatocellular carcinoma (HCC) is the leading type of primary liver cancer in humans which are linked to viral factors such as the hepatitis B virus (HBV) and the hepatitis C virus (HCV). AKT1, EGFR, HRAS, MAPK1, and PIK3R1 in HCC should have exhibited the same action as the previously described in focal adhesion (Kanehisa & Goto, 2000). Thus, VCO active components targeting these key proteins may aid in inhibiting cell proliferation and survival. However, targeting PIK3R1 specifically has been studied to suppress cell proliferation, migration, and promote apoptosis in HCC (Ai et al., 2018).

Prostate cancer (PCa) is often diagnosed among men and is the second leading cause of male cancer deaths. The targets AKT1, EGFR, and PIK3R1 are involved in the PI3K-Akt signaling pathway of PCa which lead to the inhibition of proteins necessary for apoptosis (Kanehisa & Goto, 2000). The dysregulation of apoptosis allows cancer cells to further accumulate mutations and proliferate (Pfeffer & Singh, 2018) so by having VCO active components target these proteins, apoptosis may be revived. HRAS and MAPK1 of the MAPK signaling pathway intersects with the protein involved with HSP90AA1 and indirectly leads to proliferation and survival of PCa cells. Additionally, MMP9 is involved in the process of cellular migration and invasion of PCa cells (Kanehisa & Goto, 2000). Thus, targeting these four key proteins may hinder the proliferation, survival, migration, and invasion of cancer cells, accordingly.

Colorectal cancer (CRC) occurs as a result of the accumulation of genetic alterations in tumor suppressor genes and oncogenes in the colorectal epithelium (Kanehisa & Goto, 2000). In CRC, AKT1 and PIK3R1 are involved in the PI3K-Akt signaling pathway which inhibits pro-apoptotic proteins through phosphorylation. This pathway links with the MAPK pathway of CRC where MAPK1, MAPK8, and RHOA are involved and lead to cancer cell survival and proliferation. Similarly, HRAS, MAPK1, and MAPK8 are involved in the mTOR signaling pathway that may indirectly lead to CRC cell survival and proliferation. Having VCO active components target AKT1 and PIK3R1 in CRC may incite apoptosis and targeting the rest of the key proteins may indirectly hinder the survival and proliferation of CRC cells (Kanehisa & Goto, 2000).

Fluid shear stress (FSS) is the frictional force that blood flow exerts on the blood vessel endothelial surfaces (Kanehisa & Goto, 2000) while atherosclerosis is the buildup of substances in the artery walls that restrict blood flow (Mayo Clinic, 2018). Atherosclerosis has closely aligned pathways to cancer such as cell proliferation and clonal expansion. Thus, antiproliferative therapeutic strategies may be effective in treating both atherosclerosis and cancer (Ross et al., 2001). In FSS and atherosclerosis, RHOA was shown to be linked to the cytoskeletal alignment of the cells with the blood flow while AKT1 and PIK3R1 were involved in the PI3K-AKT signaling pathway that may lead to vasodilation. Similarly, HSP90AA1 interactions may result in vasodilation. Vasodilatory proteins are anti-atherogenic. Contrarily, MAPK8 is involved in the MAPK signaling pathway of atherosclerosis. This signaling pathway is followed by matrix degeneration through MMP9. Matrix degeneration is pro-atherogenic (Kanehisa & Goto, 2000). By having VCO active components target these pro-atherogenic proteins, atherogenesis whose pathways are closely aligned with cancer may drop.

Endocrine resistance is a major hurdle in cancer treatment because endocrine therapy is essential in treating hormone-responsive breast cancer. Typically used endocrine therapy agents are selective estrogen receptor modulators (SERMs), estrogen synthesis inhibitors, and selective estrogen receptor (ER) down-regulators (Kanehisa & Goto, 2000). In endocrine resistance, MMP9 interactions lead to EGFR which is involved in the alternative activation of the PI3K-Akt and MAPK signaling pathways. AKT1 and PIK3R1 are involved in the former signaling pathway while HRAS, MAPK1, and MAPK8 are involved in the latter signaling pathway. These two signaling pathways lead to ESR1 which is involved in the alteration of the balance of co-regulatory proteins as the ER. These coregulators interact with TFs for the proliferation of transcription. This event is a mechanism behind endocrine resistance (Kanehisa & Goto, 2000). Hence, by having VCO active components target these involved key proteins, unwanted resistance for cancer treatment may be addressed.

Cellular stimuli or toxic insults activate the Phosphatidylinositol 3'-kinase (PI3K)-Akt signaling pathway which has repeated participation in the preceding pathways. This event regulates main cellular functions: transcription, translation, proliferation, growth, and survival. Phosphatidylinositol-3,4,5-triphosphate (PIP3), produced by the catalysis of PI3K, serves as a second messenger which results in the activation of AKT. AKT controls the phosphorylation of substrates involved in apoptosis, protein synthesis, metabolism, and cell cycle (Kanehisa & Goto, 2000). The interactions of AKT1, EGFR, HSP90AA1, and PIK3R1 that involve the inhibition and activation of various molecules may indirectly lead to protein synthesis, cell proliferation, angiogenesis, DNA repair,

metabolism, cell cycle progression, and cell survival in this pathway. Having VCO active components target these key proteins may result in the inhibition of the stated cell processes. In addition, by targeting EGFR and HRAS, which are part of the ErbB signaling pathway and insulin signaling pathway, and MAPK1, which is part of the MAPK signaling pathway may indirectly suppress cell proliferation, angiogenesis, and DNA repair (Kanehisa & Goto, 2000).

Signaling pathways regulating cell proliferation, survival, growth, migration, differentiation or cytoskeletal dynamism are turned on or off by Ras proteins of the Ras signaling pathway. RAS has made its appearance several times in the preceding pathways discussed as HRAS. Targeting HRAS and EGFR may suppress cellular processes including apoptosis, cell cycle arrest, cell survival, and cell growth among others. This pathway links to the PI3K-Akt signaling pathway which involves PIK3R1 and AKT1. Targeting these proteins may hinder cell survival, growth, migration, cell cycle progress, and transcription. This pathway also involves RHOA, which is linked by PIK3R1, and MAPK1 and MAPK8 which are part of the MAPK signaling pathway. Having VCO active components target RHOA and the MAPK proteins may lead to decreased cell motility and decreased gene expression, respectively (Kanehisa & Goto, 2000).

Lastly, proteoglycans contribute to the proliferation, adhesion, angiogenesis and metastasis of certain cancer types. Proteoglycans have four types namely hyaluronan (HA), chondroitin sulfate proteoglycans (CSPGs) or dermatan sulfate proteoglycans (DSPGs), keratin sulfate proteoglycans (KSPGs), and heparan sulfate proteoglycans (HSPGs). In the HA pathway; HRAS, MAPK1 and ESR1, RHOA, PIK3R1, and AKT1 interactions may indirectly cause cell growth and survival which may be hindered upon targeting. CSPG or DSPG involves AKT1, PIK3R1, EGFR, MAPK1, and MMP9. Targeting these proteins may be essential in impeding cancer cell proliferation and survival. The KSPG pathway did not involve any of the key targets while HSPG involved MAPK1 and MMP9. Their interactions may result in their phosphorylation and shedding, respectively, and indirectly lead to the inhibition of angiogenesis. Similar to HA, targeting PIK3R1, AKT1, HRAS and MAPK1 in this pathway may indirectly cause decreased cell proliferation and survival in cancer (Kanehisa & Goto, 2000).

### 3.5 Primers and Probes of Key Targets

Primers and probes are essential for the maximal efficiency of RT-PCR, a technique used for quantifying the expression of specific genes (Applied Biosystems, 2010) because they lay the foundation to allow DNA polymerase to attach new DNA nucleotides to an existing strand of nucleotides (Scitable, n.d.), and detect target sequences, respectively (Bio-Rad, n.d.). These designs serve as the groundwork for future *in vitro* studies on the anti-cancer effects of VCO. The selected



primer and probe sequences designed from the exon-exon junction sequences of the key targets to satisfy the set parameters were shown in Table 5.

Table 5. Primers and Probes of Key Targets. Table shows the key targets with their corresponding primer and probe sequences.

#	Key Target	Primer Sequence		Probe Sequence
1	AKT1	Left Primer	CCTTCCTCAGCCCTGAAG	TACTCTTTCAGACCCAGACDCG
		Right Primer	CGTTGGGTACTCCATGACA	
2	HRAS	Left Primer	ACAGGGAGCAGATCAAACGG	TGTGGAATCTCGGCAGGCTCAG
		Right Primer	GTAGGGGATGCCGTAGCTTC	
3	HSP90AA1	Left Primer	GGCTACTGATGCCCTGAGAA	CAGTACGCTTGGGAGTCTCAGCA
		Right Primer	AAAGGGCAACGTCTCAACCT	
4	MAPK1	Left Primer	TGAATTCGAAGGCTACACCA	TTTCTAACAGGCCATCTTTCAGGC
		Right Primer	CCAAAATGTGGTTCAGCTGGT	
5	EGFR	Left Primer	AGGAAATCAGCGGTTTTTGC	TGATTCAGGCTTGCCTGAAA
		Right Primer	GGTCTCAAAGGCATGGAGG	
6	MAPK8	Left Primer	ATGAAGCTCTCAACACCCG	TGTCTGGTATGATCCTTCTGAAGCAC
		Right Primer	GGATCTTGGTGGTGGAGCT	
7	RHOA	Left Primer	TCGACAGCCCTGATAGTTTGA AAA	CCAAGATGAAGCAGGA GCCGGTGA
		Right Primer	GGCACGTTGGGACAGAAATG	
8	ESR1	Left Primer	TCTTGGCAGGAACAGGGA	TAGAGGGCATGGTGGAGATCTTCGA
		Right Primer	CCGAGATGATGTA GCCAGCA	
9	PIK3R1	Left Primer	GCCATTGAAAAGAAAGTCTGGA	CAACTCTATACAGAACACAGACTCTC
		Right Primer	TGTAATTCTGCCAGGTTGCT	
10	MMP9	Left Primer	GGTGATTGACGACGCCCTTG	CTCACCTTCACTGGCGTGTACAGC
		Right Primer	CTGGATGAOGATGTCTGGT	

The set parameters were critical in designing these primers and probes to optimize RT-PCR and assure product formation (Abd-Elsalam, 2003) for good analytic performance and high-throughput gene expression assay (Bittker, 2012). Primers and probes are sequences used for *in vitro* studies such as the quantification of gene expression regarding VCO's possible TF inhibitory activity.

The information on VCO chemical constituents and their predicted targets were obtained from several public databases. Thus, the quality of the databases is reflected on the data used in this study (Supplementary information). Lack of information on the compounds may fail to generate data for some active components and targets. Moreover, when screening for the active components and targets were executed, specific screening parameters and thresholds were selected. These parameters and thresholds affect the number of active components and targets obtained for investigation. All these factors impact the final analysis. Thus, the predictive findings on the anti-cancer mechanism of VCO does not serve as hard evidence.

#### 4. CONCLUSION AND RECOMMENDATIONS

VCO's anti-cancer mechanism which is essential in determining its potential as CAM against cancer

remains unclear. Hence, the study was conducted to elucidate the anti-cancer mechanism of action that VCO depicts through an *in silico* approach.

The VCO active components determined after OB, DL, and Pa for anti-cancer activity screening were 2-heptanone and 2-pentanone which are volatile organic compounds and dihydrokaempferol, ferulic acid, and quercetin which were suggested to be TF inhibitors and are phenolic compounds. Among the cancer protein targets, AR, ESR1, ESR2, PPARD, PPARG, and RXRA were TFs and 2-pentanone had no targets. Through centrality analysis, the top ten key targets were documented as AKT1, HRAS, HSP90AA1, MAPK1, EGFR, MAPK8, RHOA, ESR1, PIK3R1, and MMP9. Subsequently, the active components were determined to be potential ligands of their corresponding key targets through docking and evaluation of their binding affinity and protein-ligand complex stability.

The top ten KEGG pathways enriched among cancer protein targets were pathways in cancer, focal adhesion, hepatocellular carcinoma, fluid shear stress and atherosclerosis, endocrine resistance, prostate cancer, colorectal cancer, PI3K-Akt signaling pathway, proteoglycans in cancer, and Ras signaling pathway. Having VCO active components targeting the key proteins involved in these specific pathways contribute to biological processes highly essential for cancer treatment and may aid in the treatment of specific cancers such as hepatocellular carcinoma, prostate cancer, and colorectal cancer. Finally, the respective primers and probes of the top ten key targets were designed.

The data used in this study was highly dependent and limited to information available in public databases. The quality of public databases, screening parameters, and thresholds used are reflected in the final analysis. Thus, the findings on the anti-cancer mechanism of VCO does not serve as hard evidence.

The group recommends the use of alternative softwares and computational tools for comparison. This approach can be used to help determine the feasibility of future studies and further *in vitro* and *in vivo* experiments. Lastly, VCO components that exhibit enzyme-based inhibition may be further investigated and have their corresponding targets designed for pertinent enzyme inhibition assays.

#### ACKNOWLEDGEMENT

We offer our sincerest gratitude to Prof. Rosario R. Rubite, PhD, and Mr. Marco C. Reyes for their guidance, insights, suggestions, and support.

#### REFERENCES

1. Abd-Elsalam, K. A. (2003). Bioinformatic tools and guideline for PCR primer design. *African Journal of Biotechnology*, 2(5), 91-95.

2. Afriza, D., Suriyah, W. H., & Ichwan, S. J. A. (2018, August). *In silico* analysis of molecular interactions between the anti-apoptotic protein survivin and dentatin, nordentatin, and quercetin. In *Journal of Physics: Conference Series* (Vol. 1073, No. 3, p. 032001). IOP Publishing.
3. Ai, X., Xiang, L., Huang, Z., Zhou, S., Zhang, S., Zhang, T., & Jiang, T. (2018). Overexpression of PIK3R1 promotes hepatocellular carcinoma progression. *Biological Research*, 51(1), 1-10.
4. Alfandary, R. (2015). 4 Tips for Efficient Primer Design. Retrieved from Genome Compiler: <http://www.genomecompiler.com/tips-for-efficient-primer-design/>
5. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402.
6. Applied Biosystems. (2010). Introduction to Gene Expression. Retrieved from <https://www.thermofisher.com/ph/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learningcenter/gene-expression-analysis-real-time-pcr-information/introduction-gene-expression.html#8>
7. Basanagouda, M., Jadhav, V., Kulkarni, M., & Rao, R. (2011). Computer Aided Prediction of Biological Activity Spectra: Study of Correlation between Predicted and Observed Activities for Coumarin-4-Acetic Acids. *Indian Journal of Pharmaceutical Sciences.*, 73(1): 88-92.
8. Bersales, L. S. G. (2018). Deaths in the Philippines, 2016. Retrieved from <https://psa.gov.ph/content/deaths-philippines-2016>
9. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2), 90.
10. Bio-Rad. (n.d.). Introduction to PCR Primer & Probe Chemistries. Retrieved from Bio-Rad Philippines: <https://www.bio-rad.com/en-ph/applications-technologies/introduction-pcr-primer-probechemistries?ID=LUSOJW3Q3>
11. Bittker, J. (2012). High-Throughput RT-PCR for small-molecule screening assays. In *Current Protocols in Chemical Biology*, 4(1): 49-63.
12. Bronowska, A.K. (2011). Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design, Thermodynamics - Interaction Studies - Solids, Liquids and Gases. IntechOpen. doi: 10.5772/19447
13. Cayman. (n.d.). 1,2-Dioleoyl-rac-glycerol. Retrieved from <https://www.caymanchem.com/product/10007863/image>
14. Chordia, N., & Kumar, A. (2018). Bioinformatics in Drug Discovery. *SciFed Journal of Protein Science*, 1 (1).
15. Craciun, D., Modra, D., Isvoran, A. (2015). ADME-Tox profiles of some food additives and pesticides. In AIP Conference Proceedings: AIP Publishing.
16. Dai, Y., Sun, L., & Qiang, W. (2018). A new strategy to uncover the anti-cancer mechanism of Chinese compound formula by integrating systems pharmacology and bioinformatics. *Evidence-Based Complementary and Alternative Medicine*, 2018.
17. Disney, A. (2020, January 2). Social network analysis 101: Centrality measures explained. Retrieved from Cambridge Intelligence: <https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/>
18. Eke, I., & Cordes, N. (2015, April). Focal adhesion signaling and therapy resistance in cancer. In *Seminars in Cancer Biology* (Vol. 31, pp. 65-75). Academic Press.
19. Filimonov, D. & Poroikov, V. (2008). Probabilistic approaches in activity prediction. In Chemoinformatics Approaches to Virtual Screening. Eds. Alexandre Varnek and Alex Tropsha. RSC Publishing, pp. 182-216.
20. Golbeck, J. (2015). Analyzing networks. In J. Golbeck, Introduction to Social Media Investigation (pp.221-235). Syngress.
21. Hashemi, M., Behrang, N., & Farahani, E. (2014). Bioinformatic analysis for anti-cancer effects of flavonoids in vegetables and fruits. In *International Conference on Biological, Civil and Environmental Engineering* (BCEE-2014).
22. Illam, S. P., Narayanankutty, A., & Raghavamenon, A. C. (2017). Polyphenols of virgin coconut oil prevent pro-oxidant mediated cell death. *Toxicology mechanisms and methods*, 27(6), 442-450.
23. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. In *Nucleic Acids Research*, 28(1), 27-30.
24. Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., & Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. In *Pharmaceutical Research*, 31(4), 1002-1014.
25. Lagorce, D., Sperandio, O., Galons, H., Miteva, M. A., & Villoutreix, B. O. (2008). FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. In *BMC Bioinformatics*, 9(1), 396.
26. Lambert, M., Jambon, S., Depauw, S., & Cordonnier, M. (2018). Targeting transcription factors for cancer treatment. In *Molecules* 23(6):1479.
27. Liao, Y., Wang, J., Jaehnig, E., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, Volume 47, Issue W1, Pages W199–W205.
28. Li, X. Y., Yang, M. D., Hu, X. Q., Cai, F. F., Chen, X. L., Chen, Q. L., & Su, S. B. (2018). Compound-target-pathway network analysis and effective mechanisms prediction of Bu-Shen-Jian-Pi formula.

- World Journal of Traditional Chinese Medicine*, 4(4), 170.
29. Lipinski, A., Lombardo F., Dominy, B., & Feeney, P. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. In *Advanced Drug Delivery Reviews*, 46:3-26.
  30. Manohar, E. C., Masa, D., & Carpiol, C. B. (2007). State of the Art: Virgin Coconut Oil Production in the Philippines. *VRGN COCONUTO*, 1.
  31. Martin, Y. C. (2005). A bioavailability score. *Journal of medicinal chemistry*, 48(9), 3164-3170.
  32. Marina, A. M., Che Man, Y. B., Nazimah, S. A. H., & Amin, I. (2009). Chemical properties of virgin coconut oil. *Journal of the American Oil Chemists' Society*, 86(4), 301-307.
  33. Mayo Clinic. (2018). Arteriosclerosis / atherosclerosis. Retrieved from <https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/symptoms-causes/syc-20350569>
  34. McLeod, S. (2019). What a p-value tells you about statistical significance. Retrieved from <https://www.simplypsychology.org/p-value.html>
  35. Morris, G., Huey, R., Lindstrom, W., Sanner, M., Belew, R., Goodsell, D., & Olson, A. (2009). Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 16: 2785-91.
  36. O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., & Hutchison, G. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3, 33.
  37. Perez, C., & Germon, R. (2016). Graph creation and analysis for linking actors: Application to social data. In R. Layton, & P. Watters, *Automating Open Source Intelligence* (pp. 103-129). Syngress.
  38. Pfeffer, C. M., & Singh, A. T. (2018). Apoptosis: a target for anti-cancer therapy. *International Journal of Molecular Sciences*, 19(2), 448.
  39. PharmMapper. (2019). Help Document. Retrieved from <http://www.lilab-ecust.cn/pharmmapper/help.html>
  40. Poroikov, V. V., Filimonov, D. A., & Associates. (n.d.). PASS User Guide. Retrieved from <http://www.pharmaexpert.ru/passonline/downloads.php>
  41. Prediger, E. (2013). How to Design Primers and Probes for PCR and qPCR. Retrieved from Integrated DNA Technologies: <https://www.idtdna.com/pages/education/decoded/article/designing-pcr-primers-and-probes>
  42. Ross, J. S., Stagliano, N. E., Donovan, M. J., Breitbart, R. E., & Ginsburg, G. S. (2001). Atherosclerosis and cancer: common molecular pathways of disease development and progression. *Annals of the New York Academy of Sciences*, 947(1), 271-293.
  43. Sadangi, C. (2015). Primer design using Primer3 software. Retrieved from [https://www.researchgate.net/publication/275713699\\_Primer\\_design\\_using\\_Primer3\\_software](https://www.researchgate.net/publication/275713699_Primer_design_using_Primer3_software)
  44. Santos, J. E. R., Villarino, B. J., Zosa, A. R., & Dayrit, F. M. (2011). Analysis of volatile organic compounds in virgin coconut oil and their sensory attributes. *Philippine Journal of Science*, 140(2), 161-171.
  45. Scitable. (n.d.). Primers. Retrieved from <https://www.nature.com/scitable/definition/primer-305/>
  46. Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498-504.
  47. Tian, S., Li, Y., Wang, J., Zhang, J., & Hou, T. (2011). ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Molecular Pharmaceutics*, 8(3), 841-851.
  48. Vandenbroucke, I. I., Vandesompele, J., Paepe, A. D., & Messiaen, L. (2001). Quantification of splice variants using real-time PCR. *Nucleic Acids Research*, 29(13), e68-e68.
  49. Veber, A., Johnson, S., Cheng, H., Smith, B., Ward, K., & Kopple, K. (2002). Molecular properties that influence the oral bioavailability of drug candidates. In *Journal of Medicinal Chemistry*, 45: 2615-2623.
  50. Verma, P., Naik, S., Nanda, P., Banerjee, S., Naik, S., & Ghosh, A. (2019). In Vitro anti-cancer Activity of Virgin Coconut Oil and its Fractions in Liver and Oral Cancer Cells. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, 19(18), 2223-2230.
  51. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). Drugbank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*, 34(suppl\_1), D668-D672.
  52. World Health Organization. (2018). Cancer. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/cancer>
  53. Yates, A., Achuthan, P., Akanni, W., Allen, J., Allen, J., Jarreta, J., . . . Plicek, P. (2020). Ensembl 2020. *Nucleic Acids Research*, Volume 48, Issue D1, Pages D682-D688.
  54. Zhang, J., & Luo, Y. (2017). Degree centrality, betweenness centrality, and closeness centrality in social network. *Advances in Intelligent Systems Research*, 300-303.

SUPPLEMENTARY INFORMATION

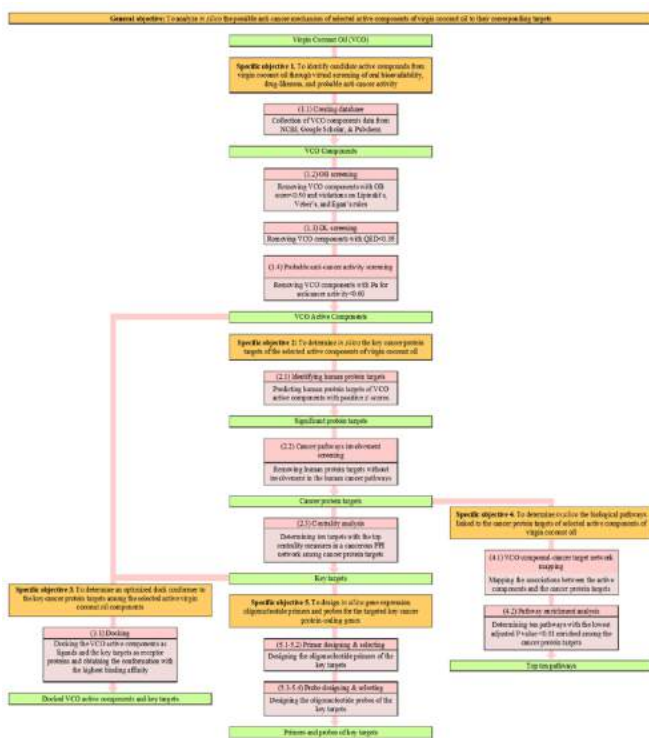


Figure S1. Outline of Specific Algorithms for Each Objective of the In Silico Study. The figure shows a detailed outline of the steps taken to satisfy the general and specific objectives of the study. The orange rectangles represent the involved objective. The green rectangles represent the input and output data while the red rectangles represent the procedure done to obtain the expected output. Lastly, the pink rectangles found beneath each procedure represent the specifications of the algorithms used.

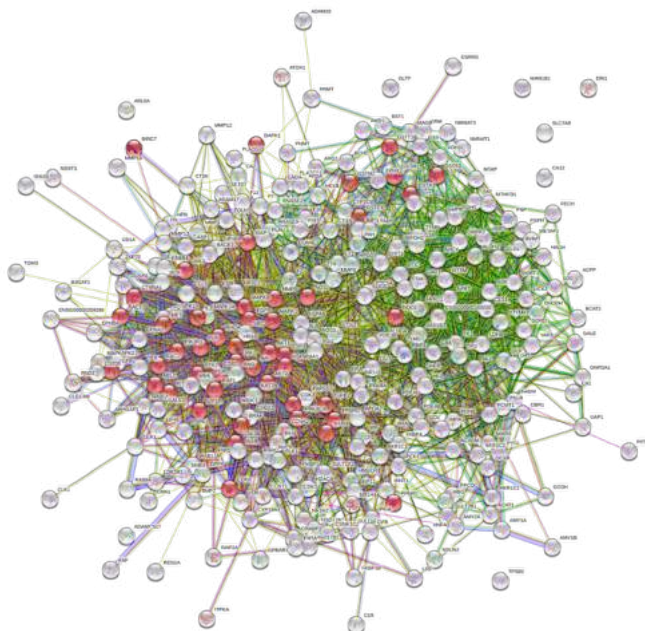


Figure S2. PPI Network of All Targets of VCO Active Components. Image shows all 291 significant protein targets of the 4 active components of VCO from STRING. The targets involved in the human pathways of cancer from the KEGG database or the cancer protein targets are represented by red nodes.

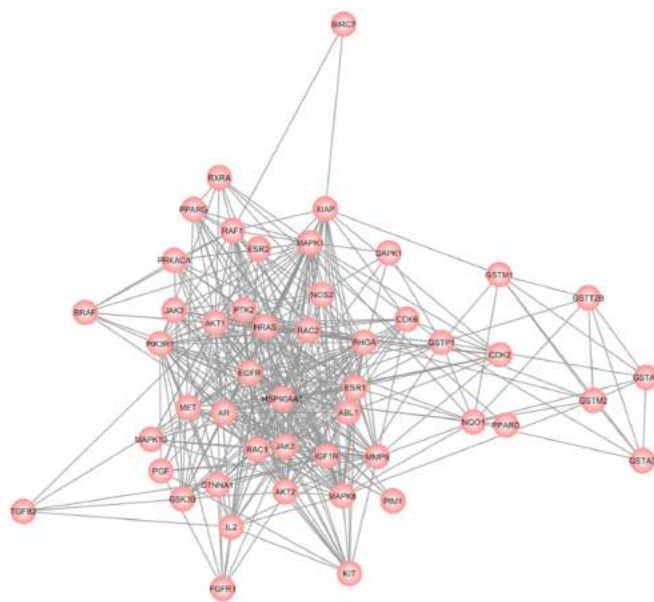

















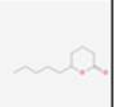
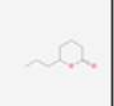
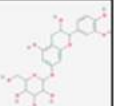
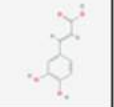
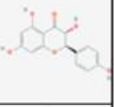
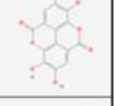
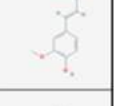
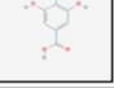


Figure S3. PPI Network of Cancer Protein Targets. Image shows the network of the protein to protein interactions (PPI) of the cancer protein targets of the VCO active components. The cancer protein targets are represented by red nodes in the cancerous network.

Table S1. VCO Components and Controls. Table shows the source, canonical SMILES, and 2D structures of the classified chemical components of VCO extracted from obtained papers and the selected controls for virtual screening.

#	VCO Component	Source	Canonical SMILES	Chemical Structure (2D)
Monoglycerides				
1	1-Monolaurin	(Dayrit et al., 2008)	<chem>CCCCCCCCCCCC(=O)OCC(O)O</chem>	
2	2-Monolaurin	(Dayrit et al., 2008)	<chem>CCCCCCCCCCCC(=O)OCC(O)CO</chem>	
Dglycerides				
3	1,2-Dipalmitin	(Dayrit et al., 2008)	<chem>CCCCCCCCCCCCCCCC(=O)OCC(O)CC(=O)OCCCCCCCCCCCCCCC</chem>	
4	1-3-Dipalmitin	(Dayrit et al., 2008)	<chem>CCCCCCCCCCCCCCCC(=O)OCC(O)C(C(=O)O)CCCCCCCCCCCCCCC</chem>	
Fatty Acids				
5	Capric Acid	(Marina et al., 2009)	<chem>CCCCCCCC(=O)O</chem>	
6	Caproic Acid	(Marina et al., 2009)	<chem>CCCCC(=O)O</chem>	

7	Caprylic Acid (Marina et al., 2009)	<chem>CCCCCCC(=O)O</chem>	
8	Lauric Acid (Marina et al., 2009)	<chem>CCCCCCCC(=O)O</chem>	
9	Linoleic Acid (Marina et al., 2009)	<chem>CCCCC=CC=CCCCC(=O)O</chem>	
10	Myristic Acid (Marina et al., 2009)	<chem>CCCCCCCC(=O)O</chem>	
11	Oleic Acid (Marina et al., 2009)	<chem>CCCCCCC=CCCCC(=O)O</chem>	
12	Palmitic Acid (Marina et al., 2009)	<chem>CCCCCCCC(=O)O</chem>	
13	Stearic Acid (Marina et al., 2009)	<chem>CCCCCCCC(=O)O</chem>	
Volatile Organic Compounds			
14	2-Heptanone (Santos et al., 2011)	<chem>CCCC(=O)C</chem>	
15	2-Pentanone (Santos et al., 2011)	<chem>CCC(=O)C</chem>	
16	Acetic Acid (Santos et al., 2011)	<chem>CC(=O)O</chem>	
17	Ethyl Acetate (Santos et al., 2011)	<chem>CCOC(=O)C</chem>	
18	Ethyl Decanoate (Santos et al., 2011)	<chem>CCCCCCCC(=O)OCC</chem>	
19	Ethyl Octanoate (Santos et al., 2011)	<chem>CCCCCCC(=O)OCC</chem>	
20	Hexanal (Santos et al., 2011)	<chem>CCCCCC=O</chem>	
21	Limonene (Santos et al., 2011)	<chem>CC1=CC(C(C1)C)=CC</chem>	
22	n-Undane (Santos et al., 2011)	<chem>CCCCCCCC</chem>	
23	Nonanal (Santos et al., 2011)	<chem>CCCCCCC=O</chem>	
24	delta-Decalactone (Santos et al., 2011)	<chem>CCCCC1CCCC(=O)O1</chem>	
25	delta-Octalactone (Santos et al., 2011)	<chem>CCCC1CCCC(=O)O1</chem>	
Phenolic Compounds			
26	7-O-Methyl Catechin (Ilam et al., 2017)	<chem>C1C(C(OC2=CC(=CC(=C21)O)OC3C(O(C(C(C3)CO)O)O)O)C4=CC(=C(C=C4)O)O)O</chem>	
27	Caffeic Acid (Marina et al., 2009)	<chem>C1=CC(=C(C=C1)C(=O)O)O</chem>	
28	Dihydrokaempferol (Ilam et al., 2017)	<chem>C1=C(C(=C(C=C1)C2C(C(=O)O)C3=C(C=C(C=C3)O)O)O)O</chem>	
29	Ellagic Acid (Ilam et al., 2017)	<chem>C1=C2C3=C(C(=C1O)O)OC(=O)C4=CC(=C(C=C4)O)O)O</chem>	
30	Ferulic Acid (Ilam et al., 2017)	<chem>CCOC(=C(C=CC(=C1)C=CC(=O)O)O)O</chem>	
31	Galic Acid (Ilam et al., 2017)	<chem>C1=C(C(=C(C=C1O)O)O)O</chem>	

32	Myricetin-3-O-glucoside	(Ilam et al., 2017)	<chem>C1=C(C=C(C(=C1O)O)O)C2=C(C=C(C=C(C=C2O)O)O)O</chem>	
33	p-Coumaric Acid	(Marina et al., 2009)	<chem>C1=CC=C(C=C1C=CC(=O)O)O</chem>	
34	Protocatechuic Acid	(Marina et al., 2009)	<chem>C1=CC=C(C=C1C(=O)O)O</chem>	
35	Quercetin	(Ilam et al., 2017)	<chem>C1=CC(=C(C=C1C2=C(C(=O)O)O)O)C=C(C=C(C=C2O)O)O</chem>	
36	Rosmarinic Acid	(Ilam et al., 2017)	<chem>C1=CC(=C(C=C1C(C(=O)O)O)O)C=C(C=C(C=C2O)O)O</chem>	
37	Syringic Acid	(Marina et al., 2009)	<chem>OCC1=CC(=CC(=C1O)O)C(=O)O</chem>	
38	Vanillic Acid	(Marina et al., 2009)	<chem>OCC1=CC(=CC(=C1O)O)O</chem>	
Sterol				
39	Cholesterol	(Dayrit et al., 2008)	<chem>CC(C)CC(C)C1CCC(C)C(C)CC3C2CC=C4C3CCC(C4)O</chem>	
Tocopherol				
40	α-Tocopherol	(Arlee et al., 2013)	<chem>CC1=C(C2=O)CCC(C2)C(C)CCC(C)CCC(C)C(C)CC(C)C(C)C</chem>	
Controls				
1	Chlorambudil (= Control)	(Wishart et al., 2006)	<chem>C1=CC(=CC=C1OCC(C(=O)O)N)C(C)CCC</chem>	
2	1,2-Dioleoyl-rac-glycerol (=Control)	(Cayman, n.d.)	<chem>CCCCCCCC=CCCCCCCCC(=O)OCC(O)CC(=O)CCCCCCCC=CCCCCCCCC(=O)O</chem>	
43	FDH6 (= Control)	(Lambert et al., 2018)	<chem>C1=CC=C(C=C1C2=NC3=C(C=C2)C(F)F)C=C(C=C3)C(=O)N(C4=CC=CC=C4)F</chem>	
44	Etiopidine (= Control)	(Lambert et al., 2018)	<chem>CC1=CC2=C(N=CC2=C1C3=C1NC4=CC=CC=C43)C</chem>	

#### List of References for Supplementary Table S1

- Arlee, R., Suanphairoch, S., & Pakdeechuan, P. (2013). Differences in chemical components and antioxidant-related substances in virgin coconut oil from coconut hybrids and their parents. *International Food Research Journal*, 20(5), 2103.
- Cayman. (n.d.). 1,2-Dioleoyl-rac-glycerol. Retrieved from <https://www.caymanchem.com/product/10007863/image>
- Dayrit, F. M., Buenafe, O. E. M., Chainani, E. T., & De Vera, I. M. S. (2008). Analysis of monoglycerides, diglycerides, sterols, and free fatty acids in coconut

(*Cocos nucifera* L.) oil by 31P NMR spectroscopy. *Journal of agricultural and food chemistry*, 56(14), 5765-5769.

- Ilam, S. P., Narayanankutty, A., & Raghavamenon, A. C. (2017). Polyphenols of virgin coconut oil prevent oxidant mediated cell death. *Toxicology mechanisms and methods*, 27(6), 442-450.
- Lambert, M., Jambon, S., Depauw, S., & Cordonnier, M. (2018). Targeting transcription factors for cancer treatment. *In Molecules* 23(6):1479.
- Marina, A. M., Che Man, Y. B., Nazimah, S. A. H., & Amin, I. (2009). Chemical properties of virgin coconut oil. *Journal of the American Oil Chemists' Society*, 86(4), 301-307.
- Santos, J. E. R., Villarino, B. J., Zosa, A. R., & Dayrit, F. M. (2011). Analysis of volatile organic compounds in virgin coconut oil and their sensory attributes. *Philippine Journal of Science*, 140(2), 161-171.
- Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl\_1), D668-D672.

Table S2. Oral Bioavailability of VCO Components and Controls. Table shows the oral bioavailability and filter results of the components of VCO and the used controls. In this table, 1 signifies that the compound passed through the specific filter by satisfying its rule while 0 signifies a violation was made by the compound. Components that generated an OB score > 0.5 and passed the three filters were selected for further screening.

#	VCO Component	OB score	Lipinski's	Veber's	Egan's	Orally Bioavailable
1	1-Monolaurin	0.55	1	0	1	NO
2	2-Monolaurin	0.55	1	0	1	NO
3	1,2-Dipalmitin	0.17	0	0	0	NO
4	1,3-Dipalmitin	0.17	0	0	0	NO
5	Capric Acid	0.58	1	1	1	YES
6	Caproic Acid	0.58	1	1	1	YES
7	Caprylic Acid	0.58	1	1	1	YES
8	Lauric Acid	0.58	1	1	1	YES
9	Linoleic Acid	0.58	0	0	0	NO
10	Myristic Acid	0.58	1	0	1	NO
11	Oleic Acid	0.58	0	0	0	NO
12	Palmitic Acid	0.58	0	0	1	NO
13	Stearic Acid	0.58	0	0	0	NO
14	2-Heptanone	0.55	1	1	1	YES
15	2-Pentanone	0.55	1	1	1	YES
16	Acetic Acid	0.58	1	1	1	YES
17	Ethyl Acetate	0.55	1	1	1	YES
18	Ethyl Decanoate	0.55	1	1	1	YES
19	Ethyl Octanoate	0.55	1	1	1	YES
20	Hexanal	0.55	1	1	1	YES
21	Limonene	0.55	1	1	1	YES
22	n-Octane	0.55	0	1	1	NO
23	Nonanal	0.55	1	1	1	YES
24	5-Decalactone	0.55	1	1	1	YES
25	5-Octalactone	0.55	1	1	1	YES

#	VCO Component	WQED	UWGED	DRUG-LIKE		
26	7-O-Methyl Catechin	0.17	0	0	0	NO
27	Caffeic Acid	0.56	1	1	1	YES
28	Dihydrokaempferol	0.55	1	1	1	YES
29	Ellagic Acid	0.55	1	0	0	NO
30	Ferulic Acid	0.56	1	1	1	YES
31	Gallic Acid	0.56	1	0	1	NO
32	Myricetin-3-O-glucoside	0.17	0	0	0	NO
33	p-Coumaric Acid	0.56	1	1	1	YES
34	Protocatechuic Acid	0.56	1	1	1	YES
35	Quercetin	0.55	1	1	1	YES
36	Rosmarinic Acid	0.56	1	0	0	NO
37	Syringic Acid	0.56	1	1	1	YES
38	Vanillic Acid	0.56	1	1	1	YES
39	Cholesterol	0.55	0	1	0	NO
40	α-Tocopherol	0.55	0	0	0	NO
41	Chlorambucil (+ control)	0.56	1	1	1	YES
42	1,2-Dioleoyl-rac-glycerol (- control)	0.17	0	0	0	NO

Table S3. Drug-Likeness of OB-Screened VCO Components and Controls. Table shows the drug-likeness scores (Weighted QED and Unweighted QED) of the VCO components and used controls that passed OB-screening. Components generated an W QED and UW QED score > 0.35 were selected for further screening.

#	VCO Component	WQED	UWGED	DRUG-LIKE
1	Capric Acid	0.501	0.528	YES
2	Caproic Acid	0.565	0.603	YES
3	Caprylic Acid	0.549	0.579	YES
4	Lauric Acid	0.443	0.467	YES
5	2-Heptanone	0.507	0.446	YES
6	2-Pentanone	0.496	0.441	YES
7	Acetic Acid	0.471	0.527	YES
8	Ethyl Acetate	0.445	0.503	YES
9	Ethyl Decanoate	0.346	0.422	NO
10	Ethyl Octanoate	0.378	0.46	YES
11	Hexanal	0.441	0.415	YES
12	Limonene	0.488	0.328	NO
13	Nonanal	0.411	0.383	YES
14	δ-Decalolone	0.399	0.469	YES
15	δ-Octalolone	0.476	0.519	YES
16	Caffeic Acid	0.561	0.655	YES
17	Dihydrokaempferol	0.639	0.638	YES
18	Ferulic Acid	0.748	0.802	YES
19	p-Coumaric Acid	0.677	0.749	YES
20	Protocatechuic Acid	0.554	0.632	YES
21	Quercetin	0.458	0.487	YES
22	Syringic Acid	0.767	0.791	YES
23	Vanillic Acid	0.704	0.759	YES
24	Chlorambucil (+ control)	0.766	0.802	YES
25	1,2-Dioleoyl-rac-glycerol (- control)	0.000	0.000	NO

Table S4. Probability of Anti-Cancer and TF Inhibitory Activity of OB- and DL-Screened VCO Components and Controls. Table shows the anti-cancer activity probability scores (Pa and Pi) of the VCO components and used controls that passed both OB- and DL-screening. Components that generated a Pa value > 0.6 for anticarcinogenic and antineoplastic activities were suggested to exhibit anti-cancer activity. Additionally, the probability scores for TF inhibitory activity of the candidate anti-cancer VCO components and the used controls were shown. The active components that generated a Pa value > 0.4 for inhibitory activity were suggested to exhibit the mentioned activity. The “\*” represents cells that were intended to be empty.

#	VCO Component	Anticarcinogenic Activity		Antineoplastic Activity		Anti-cancer Activity	TF Inhibition		Inhibitory Activity
		Pa	Pi	Pa	Pi		Pa	Pi	
1	Capric Acid	0.359	0.039	0.460	0.058	NO	*	*	*
2	Caproic Acid	0.359	0.039	0.460	0.058	NO	*	*	*
3	Caprylic Acid	0.359	0.039	0.460	0.058	NO	*	*	*
4	Lauric Acid	0.359	0.039	0.460	0.058	NO	*	*	*
5	2-Heptanone	0.261	0.077	0.696	0.004	YES	0.332	0.017	NO
6	2-Pentanone	0.299	0.078	0.715	0.004	YES	0.347	0.012	NO
7	Acetic Acid	No data (Carbons <3)					*	*	*
8	Ethyl Acetate	0.315	0.053	0.468	0.053	NO	*	*	*
9	Ethyl Octanoate	0.301	0.058	0.455	0.061	NO	*	*	*
10	Hexanal	0.241	0.087	0.278	0.012	NO	*	*	*
11	Nonanal	0.241	0.087	0.278	0.012	NO	*	*	*
12	δ-Decalolone	0.217	0.104	0.443	0.07	NO	*	*	*
13	δ-Octalolone	0.209	0.11	0.442	0.069	NO	*	*	*
14	Caffeic Acid	0.571	0.014	0.530	0.063	NO	*	*	*
15	Dihydrokaempferol	0.792	0.005	0.715	0.024	YES	0.417	0.006	YES
16	Ferulic Acid	0.616	0.012	0.601	0.045	YES	0.417	0.007	YES
17	p-Coumaric Acid	0.559	0.015	0.52	0.065	NO	*	*	*
18	Protocatechuic Acid	0.387	0.033	0.407	0.101	NO	*	*	*
19	Quercetin	0.757	0.007	0.797	0.012	YES	0.563	0.006	YES
20	Syringic Acid	0.413	0.029	0.430	0.094	NO	*	*	*
21	Vanillic Acid	0.413	0.029	0.368	0.119	NO	*	*	*

Table S5. Protein Targets of VCO Active Components. Table shows the predicted significant protein target of each active component of VCO which totals 291 distinct proteins. The proteins are represented by their HGNC symbols. The second column shows the targets of 2-heptanone; the third column shows the targets of 2-pentanone; the fourth column indicates the targets of dihydrokaempferol; the fifth column indicates the targets of ferulic acid; and the targets of quercetin are indicated at the sixth column. The “\*” represents cells that were intended to be empty.

#	Significant Targets				
	2-Heptanone	2-Pentanone (No Significant Targets)	Dihydrokaempferol	Ferulic Acid	Quercetin
1	ABL1	*	ABL1	ABO	ABL1
2	ACADM	*	ABO	ACAT1	ABO
3	ADAMTS17	*	ADAM17	AK1	ACAT1
4	ADH1C	*	ADAMTS17	AKT1	ACPP
5	AKR1B1	*	ADH1C	AKT2	ADAM17
6	AKR1C2	*	ADH5	ALB	ADAM33
7	ALB	*	ADK	AMD1	ADAMTS17
8	AR	*	AHCY	AMY1B	ADH5
9	BIRC7	*	AKR1B1	APRT	ADK
10	BRAF	*	AKR1C1	ARG2	AHCY
11	CA2	*	AKR1C3	ARHGAP1	AKR1B1
12	CCNA2	*	AKT1	BACE1	AKR1C1
13	CDK2	*	ALB	BAT1	AKR1C2
14	CHEK1	*	ALDH2	BCAT2	AKR1C3
15	CRABP2	*	ALDOA	BMP7	AKT2
16	DCK	*	AMD1	BRAF	ALDH2
17	DHODH	*	AMY1B	BST1	ALDOA
18	DPP4	*	AMY2A	CBS	AMD1
19	EGFR	*	ANG	CCL5	AMY1A
20	EPHB4	*	APOA2	CDK2	AMY1A
21	ERBB4	*	AR	CDK6	ANG
22	F10	*	ARHGAP1	CDK7	APRT
23	F2	*	ARL5A	CFB	ARG2
24	FKBP1A	*	ATOX1	CHEK1	ARHGAP1

25	FOLH1	*	B3GAT1	CLK1	ARL5A
26	GSK3B	*	BACE1	CMA1	ARSA
27	GSTP1	*	BCAT2	CTSK	ATIC
28	HCK	*	BCHE	CTSS	ATOX1
29	HDAC8	*	BHMT	DAFK1	B3GAT1
30	HNF4G	*	BIRC7	DKC	BACE1
31	HNMT	*	BRAF	DCXR	BCAT2
32	HPN	*	BST1	DHFR	BCEH
33	HSD11B1	*	CA1	DTYMK	BHMT
34	HSD17B1	*	CA2	DUT	BIRC7
35	INSR	*	CASP1	EGFR	BMP7
36	JAK2	*	CBR1	EIF4E	BST1
37	KDR	*	CBS	EPHA2	C1R
38	KIT	*	CCNA2	EPHB4	CA12
39	LCK	*	CCNT1	ESR1	CA2
40	LGALS3	*	CD1A	F10	CASP1
41	LTA4H	*	CD209	F7	CBR1
42	MAOB	*	CDA	FGFR1	CBS
43	MAPK1	*	CDK2	FGG	CCNT1
44	MAPK10	*	CDK5R1	FKBP1A	CD1A
45	MAPK14	*	CDK8	FKBP1B	CD209
46	MAPK8	*	CES1	GALE	CDK2
47	MET	*	CFB	GCK	CDK8
48	METAP2	*	CFD	GLO1	CDK7
49	MTAP	*	CHEK1	GMFR	CES1
50	NOS3	*	CHIT1	GMFR2	CFB
51	NQO1	*	CLEC4M	GP1BA	CHT1
52	PDE3B	*	CLK1	GSK3B	CLEC4M
53	PDE4B	*	CRABP1	GSS	CLK1
54	PDPK1	*	CSNK1G2	GSTA1	CSNK1G2
55	PM1	*	CTNNA1	GSTA3	CTSK
56	PKIA	*	CTSK	GSTM1	CTSS
57	PLA2G2A	*	CTSS	GSTM2	CYP2C9
58	PNMT	*	CYP19A1	GSTT2B	DKC
59	PPARG	*	CYP2C9	HADH	DHODH
60	PPIA	*	DKC	HDAC8	DPP4
61	PTPN1	*	DHFR	HMGCR	DTYMK
62	RARG	*	DHODH	HPGDS	EGFR
63	SETD7	*	DPP4	HRAS	EIF4E
64	SHBG	*	DUSP8	HSP90AA1	ELANE
65	SORD	*	DUT	HSPA8	EPHA2
66	SULT2B1	*	EEA1	IMPDH1	EPHB4
67	TEK	*	EGFR	ISG20	EPHX2
68	TTR	*	EIF4E	ITPKA	ERI1
69	TYMS	*	ELANE	KAT2B	ESR1
70	YARS	*	EPHB4	KDR	ESR2
71	*	*	EPHX2	KIF11	ESRRG
72	*	*	ERBB4	KIT	F10
73	*	*	ESR1	LYZ	F11
74	*	*	ESR2	MAN1B1	F2
75	*	*	ESRRG	MAPK10	FABP4
76	*	*	F10	MAPK14	FAP
77	*	*	F11	MAPK8	FGFR1
78	*	*	F2	MIF	FHIT
79	*	*	F7	MME	FKBP1A
80	*	*	FABP4	MMP16	GBA
81	*	*	FECH	MMR8	GCK
82	*	*	FGG	MTAP	GLO1
83	*	*	FKBP1A	NDST1	GLTP
84	*	*	FKBP1B	NGAL	GM2A
85	*	*	FNTA	NOS2	GMFR
86	*	*	GART	NOS3	GMFR2
87	*	*	GBA	OTC	GP1BA
88	*	*	GCDH	PAH	GPBAR1
89	*	*	GCK	PAK6	GPI
90	*	*	GMFR	PAPSS1	GSK3B
91	*	*	GNPDA1	POK1	GSR
92	*	*	GPI	PDE4B	GSTA3
93	*	*	GSK3B	PDE4D	GSTM2
94	*	*	GSR	PDE5A	GSTT2B
95	*	*	GSTA3	PDPK1	HAGH
96	*	*	GSTM1	PIM1	HCK
97	*	*	GSTM2	PKLR	HDAC8
98	*	*	GSTP1	PSPH	HEXB
99	*	*	GSTT2B	PTK2	HINT1
100	*	*	HCK	RAB11A	HK1

101	*	*	HDAC8	RAC1	HMGCR
102	*	*	HEXB	RAC2	HPRT1
103	*	*	HK1	RAF1	HRAS
104	*	*	HMGCR	RAN	HSD17B1
105	*	*	HNMT	RAP2A	HSP90AA1
106	*	*	HPGDS	RARG	IL2
107	*	*	HRAS	REN	IMPDH2
108	*	*	HSD11B1	RHOA	INSR
109	*	*	HSD17B1	RNASE2	JAK2
110	*	*	IGF1R	SDS	JAK3
111	*	*	IL2	SHMT1	KAT2B
112	*	*	IMPA1	SOD2	KIF11
113	*	*	IMPDH2	SULT1E1	KIT
114	*	*	INSR	SULT2B1	LCK
115	*	*	ITK	TGM3	LCN2
116	*	*	ITPKA	TK1	LDHB
117	*	*	JAK3	TP1	LGALS3
118	*	*	KAT2B	TPSB2	LSS
119	*	*	KDR	UAP1	LTA4H
120	*	*	KIF11	UCK2	LYZ
121	*	*	KIT	UMPS	MAN1B1
122	*	*	LCK	WARS	MAPK10
123	*	*	LGALS3	YARS	MAPK14
124	*	*	LTA4H	*	MAPK8
125	*	*	LYZ	*	MAPKAPK2
126	*	*	MAPK10	*	MMP12
127	*	*	MAPK14	*	MTAP
128	*	*	MAPK8	*	MTFSD1
129	*	*	MAPKAPK2	*	NDST1
130	*	*	MET	*	NMNAT1
131	*	*	METAP2	*	NMNAT3
132	*	*	MIF	*	NOS2
133	*	*	MME	*	NQO1
134	*	*	MMP12	*	NR1Q2
135	*	*	MMP13	*	NR3C1
136	*	*	MMP3	*	NR3C2
137	*	*	MMP8	*	NT5M
138	*	*	MMP9	*	OTC
139	*	*	NMNAT1	*	PAH
140	*	*	NONE	*	PCK1
141	*	*	NOS2	*	PCMT1
142	*	*	NOS3	*	PDE3B
143	*	*	NQO1	*	PDE4B
144	*	*	NR1H2	*	PDE4D
145	*	*	NR1H3	*	PDE5A
146	*	*	NR1H4	*	PDHB
147	*	*	NR1Q2	*	PDPK1
148	*	*	NSUN2	*	PGR
149	*	*	OAT	*	PK3R1
150	*	*	OTC	*	PM1
151	*	*	PAH	*	PITPNA
152	*	*	PCK1	*	PLA2G10
153	*	*	PDE4D	*	PLAU
154	*	*	PDE5A	*	PLK1
155	*	*	PDHB	*	PPARA
156	*	*	PDPK1	*	PPCC
157	*	*	PGF	*	PPIA
158	*	*	PGR	*	PRKACA
159	*	*	PIM1	*	PTPN1
160	*	*	PKIA	*	PYGL
161	*	*	PLAU	*	RAB11A
162	*	*	PNP	*	RAB5A
163	*	*	PPARA	*	RAB9A
164	*	*	PPARD	*	RAN
165	*	*	PPIA	*	REG1A
166	*	*	PRKACA	*	RHEB
167	*	*	PYGL	*	RHOA
168	*	*	RBP4	*	RND3
169	*	*	RHEB	*	SELE
170	*	*	RNASE3	*	SHBG
171	*	*	RXRA	*	SHMT1
172	*	*	SELE	*	SOD2
173	*	*	SETD7	*	SORD
174	*	*	SHBG	*	SRC
175	*	*	SHMT1	*	SRM



176	*	*	SOD2	*	SULT2B1
177	*	*	SORD	*	TK1
178	*	*	SULT1E1	*	TNK2
179	*	*	SULT2A1	*	TPSB2
180	*	*	TGFB2	*	TYMP
181	*	*	TGM3	*	TYMS
182	*	*	TPSB2	*	UCK2
183	*	*	TYMP	*	UMPS
184	*	*	TYMS	*	VDR
185	*	*	UCK2	*	WARS
186	*	*	UMPS	*	XIAP
187	*	*	VDR	*	YARS
188	*	*	WARS	*	ZAP70
189	*	*	XIAP	*	-
190	*	*	YARS	*	-
191	*	*	ZAP70	*	-

Table S6. Cancer Protein Targets of VCO Active Components. Table shows the 49 distinct cancer protein targets of the VCO active components and their corresponding HGNC symbols.

#	Targets	Symbol
1	Proto-oncogene tyrosine-protein kinase ABL1	ABL1
2	RAC-alpha serine/threonine-protein kinase	AKT1
3	RAC-beta serine/threonine-protein kinase	AKT2
4	Androgen receptor	AR
5	Baculoviral IAP repeat-containing protein 7	BIRC7
6	B-Raf proto-oncogene serine/threonine-protein kinase	BRAF
7	Cell division protein kinase 2	CDK2
8	Cell division protein kinase 6	CDK6
9	Catenin alpha-1	CTNNA1
10	Death-associated protein kinase 1	DAPK1
11	Epidermal growth factor receptor	EGFR
12	Estrogen receptor	ESR1
13	Estrogen receptor beta	ESR2
14	Basic fibroblast growth factor receptor 1	FGFR1
15	Glycogen synthase kinase-3 beta	GSK3B
16	Glutathione S-transferase A1	GSTA1
17	Glutathione S-transferase A3	GSTA3
18	Glutathione S-transferase Mu 1	GSTM1
19	Glutathione S-transferase Mu 2	GSTM2
20	Glutathione S-transferase P	GSTP1
21	Glutathione S-transferase theta-2	GSTT2B
22	GTPase HRas	HRAS
23	Heat shock protein HSP 90-alpha	HSP90AA1
24	Insulin-like growth factor 1 receptor	IGF1R
25	Interleukin-2	IL2
26	Tyrosine-protein kinase JAK2	JAK2
27	Tyrosine-protein kinase JAK3	JAK3
28	Mast/stem cell growth factor receptor	KIT
29	Mitogen-activated protein kinase 1	MAPK1
30	Mitogen-activated protein kinase 10	MAPK10
31	Mitogen-activated protein kinase 8	MAPK8
32	Hepatocyte growth factor receptor	MET
33	Matrix metalloproteinase-9	MMP9
34	Nitric oxide synthase, inducible	NOS2
35	NAD(P)H dehydrogenase [quinone] 1	NQO1
36	Placenta growth factor	PGF
37	Phosphatidylinositol 3-kinase regulatory subunit alpha	PIK3R1
38	Proto-oncogene serine/threonine-protein kinase Pim-1	PIM1
39	Peroxisome proliferator-activated receptor delta	PPARD
40	Peroxisome proliferator-activated receptor gamma	PPARG
41	cAMP-dependent protein kinase, alpha-catalytic subunit	PRKACA
42	Focal adhesion kinase 1	PTK2
43	Ras-related C3 botulinum toxin substrate 1	RAC1
44	Ras-related C3 botulinum toxin substrate 2	RAC2
45	RAF proto-oncogene serine/threonine-protein kinase	RAF1
46	Transforming protein RhoA	RHOA
47	Retinoic acid receptor RXR-alpha	RXRA
48	Transforming growth factor beta-2	TGFB2
49	Baculoviral IAP repeat-containing protein 4	XIAP

Table S7. Topological Data of Cancerous PPI Network. Table shows the topological features of the PPI network of cancer protein targets that were generated after network topological analysis.

Parameters	Numerical Value
Clustering coefficient	0.673
Connected components	1
Network diameter	3
Network radius	2
Network centralization	0.4
Shortest paths	2352 (100%)
Characteristic path length	1.8
Avg. number of neighbors	15.592
Number of nodes	49
Number of edges	382
Network density	0.325
Network heterogeneity	0.578
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	0

Table S8. Binding Affinity of the Docked VCO Active Components and Key Targets. Table shows all nine binding affinities of the different conformations of the docked VCO active components (ligands) with their corresponding key cancer protein targets (receptor proteins) and Figure S3 reference.

#	Ligand	Receptor	Binding Affinity (kcal/mol)								
			1st	2nd	3rd	4th	5th	6th	7th	8th	9th
1	2-Heptanone	MAPK1	-4.1	-3.9	-3.9	-3.9	-3.9	-3.9	-3.8	-3.7	-3.7
		EGFR	-3.7	-3.6	-3.6	-3.5	-3.5	-3.4	-3.4	-3.4	-3.4
		MAPK8	-4.4	-4.4	-4.3	-4.0	-4.0	-3.9	-3.8	-3.7	-3.7
2	Dihydrokaempferol	AKT1	-8.0	-7.3	-7.3	-7.2	-7.0	-6.9	-6.8	-6.8	-6.7
		HRAS	-6.5	-6.4	-6.2	-6.1	-6.1	-5.8	-5.8	-5.7	-5.6
		EGFR	-6.8	-6.8	-6.4	-6.4	-6.2	-6.1	-6.1	-6.1	-6.0
		MAPK8	-8.1	-8.0	-7.5	-7.4	-7.3	-7.2	-6.9	-6.7	-6.6
		ESR1	-7.5	-6.9	-6.7	-6.3	-6.2	-6.1	-6.1	-6.0	-6.0
3	Ferulic Acid	MMP9	-8.3	-8.0	-8.0	-7.8	-7.6	-7.5	-7.4	-7.2	-7.1
		AKT1	-6.6	-6.2	-6.0	-5.9	-5.8	-5.7	-5.6	-5.5	-5.4
		HRAS	-5.3	-5.3	-5.2	-5.1	-5.0	-5.0	-5.0	-4.9	-4.9
		HSP90AA1	-6.8	-6.6	-6.4	-5.9	-5.7	-5.7	-5.6	-5.6	-5.4
		EGFR	-5.8	-5.5	-5.4	-5.2	-5.2	-5.2	-5.2	-5.1	-5.0
		MAPK8	-6.7	-6.2	-6.1	-6.0	-5.9	-5.9	-5.9	-5.9	-5.8
		RHOA	-5.9	-5.6	-5.5	-5.4	-5.3	-5.3	-5.2	-5.2	-5.1
4	Quercetin	ESR1	-6.3	-6.0	-5.5	-5.4	-5.4	-5.3	-5.1	-5.0	-5.0
		HRAS	-7.5	-7.1	-6.7	-6.5	-6.3	-6.2	-6.2	-6.1	-6.0
		HSP90AA1	-8.0	-7.7	-7.4	-7.2	-7.1	-7.0	-6.9	-6.8	-6.7
		EGFR	-7.3	-7.2	-7.1	-6.9	-6.8	-6.8	-6.6	-6.6	-6.5
		MAPK8	-8.4	-7.9	-7.6	-7.5	-7.5	-7.5	-7.3	-7.1	-6.9
		RHOA	-6.9	-6.6	-6.4	-6.4	-6.2	-6.2	-6.1	-6.0	-6.0

Table S9. Primers of Key Targets. Table shows the left and right primers of the key cancer protein targets with their corresponding starting position, number of base pairs (length), melting temperature (Tm), guanine-cytosine (GC) content, query cover (QC), E value, and the product size of the amplicon.

#	Key Target	Primer Parameters			
1	AKT1	Left Primer	CCTTCCTCACGCCCTGAAG	Start: 56	Length: 20 bp
				Tm: 60.0 C	GC: 60.0 %
		Right Primer	CGTTGGCGTACTCCATGACA	Start: 127	Length: 20 bp
				Tm: 60.4 C	GC: 55.0 %
		Amplicon	Product Size: 72 bp	QC: 100 %	E value: 0.47
2	HRAS	Left Primer	ACAGGGAGCAGATCAAACGG	Start: 176	Length: 20 bp
				Tm: 60.0 C	GC: 55.0 %
		Right Primer	GTAGGGGATGCGCTAGCTTC	Start: 312	Length: 20 bp
				Tm: 60.0 C	GC: 60.0 %
		Amplicon	Product Size: 137 bp	QC: 100 %	E value: 0.47
3	HSP90AA1	Left Primer	GGCTACTGATGCCTGAGGAA	Start: 204	Length: 20 bp
				Tm: 59.2 C	GC: 55.0 %
		Right Primer	AAAGGGCAAGCTCAAACCT	Start: 277	Length: 20 bp
				Tm: 59.9 C	GC: 50.0 %
		Amplicon	Product Size: 74 bp	QC: 100 %	E value: 0.12
4	MAPK1	Left Primer	TGAATCCAAGGGCTACACCA	Start: 107	Length: 21 bp
				Tm: 59.3 C	GC: 47.6 %
		Right Primer	CCAAAATGTGGTTCAAGCTGGT	Start: 232	Length: 21 bp
				Tm: 59.3 C	GC: 47.6 %
		Amplicon	Product Size: 126 bp	QC: 100 %	E value: 0.47

5	EGFR	Left Primer	AGGAAATCACAGGGTTTTTGC	Start: 63	Length: 21 bp
				Tm: 57.2 C	GC: 42.9 %
		Right Primer	GGTTCTCAAAGGCATGGAGG	Start: 134	Length: 20 bp
		Amplicon	Product Size: 72 bp	Tm: 58.5 C	GC: 55.0 %
				QC: 100%	E value: 0.47
6	MAPK3	Left Primer	ATGAAGCTCTCCAACACCCG	Start: 67	Length: 20 bp
				Tm: 60.0 C	GC: 55.0 %
		Right Primer	GGATCTTTGGTGGAGCT	Start: 141	Length: 20 bp
		Amplicon	Product Size: 75 bp	Tm: 59.7 C	GC: 55.0 %
				QC: 100%	E value: 0.47
7	RHOA	Left Primer	TGACAGCCCTGATAGTTTAGAAAA	Start: 101	Length: 25 bp
				Tm: 60.0 C	GC: 40.0 %
		Right Primer	GGCACGTTGGGACAGAAATG	Start: 178	Length: 20 bp
		Amplicon	Product Size: 76 bp	Tm: 59.8 C	GC: 55.0 %
				QC: 100%	E value: 0.47
8	ESR1	Left Primer	TCTGGACAGGAACAGGGA	Start: 130	Length: 20 bp
				Tm: 60.1 C	GC: 55.0 %
		Right Primer	CCGAGATGATGAGCCAGCA	Start: 206	Length: 20 bp
		Amplicon	Product Size: 77 bp	Tm: 59.8 C	GC: 55.0 %
				QC: 100%	E value: 0.47
9	PIK3R1	Left Primer	GCCATTGAAAAGAAAGTCTGGA	Start: 78	Length: 23 bp
				Tm: 59.4 C	GC: 43.5 %
		Right Primer	TGTAATTCTGCCAGTTGCT	Start: 152	Length: 21 bp
		Amplicon	Product Size: 75 bp	Tm: 60.0 C	GC: 47.6 %
				QC: 100%	E value: 0.011
10	MMP9	Left Primer	GGTGATTGACGACGCCITG	Start: 34	Length: 20 bp
				Tm: 59.8 C	GC: 55.0 %
		Right Primer	CTGGATGACGATGTCTGOGT	Start: 136	Length: 20 bp
		Amplicon	Product Size: 103 bp	Tm: 60.2 C	GC: 55.0 %
				QC: 100%	E value: 0.47

Table S10. Probes of Key Targets. Table shows the probe sequences of the key targets with their length, melting temperature (Tm), annealing temperature (Ta), guanine-cytosine (GC) content, query cover (QC), change in Gibbs free energy ( $\Delta G$ ), and amplicon length.

#	Key Target	Probe Sequence	Length	Tm (°C)	Ta (°C)	GC (%)	QC (%)	$\Delta G$ (kcal/mole)	Amplicon Length
1	AKT1	TACTCTTTCCAGACCACGACCGC	24	69	55	58	100	0.47	72
2	HRAS	TGTGGAATCTCGGCAGCCTCAG	22	68	55	59	100	0.99	137
3	HSP90AA1	CAGTACGCTTGGAGTCCCTCAGCA	24	67	56	58	100	-1.66	74
4	MAPK1	TTTCTAACAGGCCCACTTTCCAGGG	26	68	54	50	100	-1.16	126
5	EGFR	TGATTCAGGCTTGCCGTAATA	21	65	52	48	100	-5.06	72
6	MAPK3	TGCTGGTATGATCCTCTGAAGCAG	26	66	55	46	100	-0.15	75
7	RHOA	CCAAGATGAAGCAGGAGCCGGTGA	24	67	55	58	100	-0.68	76
8	ESR1	TAGAGGGCATGGTGGAGATCTTCGA	25	68	54	52	100	0.64	77
9	PIK3R1	CAACTCTATACAGAACACAGAGCTCC	26	64	54	46	100	-0.97	75
10	MMP9	CTCACCTTCACTCGCGTGTACAGC	24	68	54	58	100	-1.00	103

# Preliminary Evaluation of *rbcL*, *matK*, and SRAP Markers for the Molecular Characterization of Five Philippine *Allium sativum* varieties

Bigtas, Allisandra Isabel B.<sup>1</sup>, Moreno, Patrick Gabriel G.<sup>1</sup>, and Heralde, Francisco III M.<sup>1</sup>

<sup>1</sup>Department of Biochemistry and Molecular Biology, College of Medicine, University of the Philippines Manila, Pedro Gil St., Ermita, Manila

## Email address:

saab.bigtas@gmail.com, pgmoreno@gmail.com

## To cite:

Bigtas AIB, Moreno Patrick GG., and Heralde FIII M. 2020. Preliminary Evaluation of *rbcL*, *matK*, and SRAP Markers for the Molecular Characterization of Five Philippine *Allium sativum* varieties. PJBMB. Vol. I, No. 1, 2020, pp. 54-60. doi: 10.5555/pjbmb.ph.2020.01.01.54

Received: 07 25, 2019; Accepted: 12 18, 2020; Published: 01 07, 2021

**Abstract:** *Allium sativum*, commonly known as “*bawang*”, is one of the economically important crops in the Philippines mainly for spice and other uses including herbal medicine for hypertension and cancer. Several studies showing therapeutic properties of garlic often do not disclose the variety used while typical varietal identification rely on phenotypic or agronomic traits which may be inadequate to define distinct variants. This study aimed to evaluate the utility of chloroplast genes, *rbcL* and *matK*, as well as Sequence-Related Amplified Polymorphism (SRAP) to define and characterize five *A. sativum* varieties grown in the Philippines. Five garlic varieties (i.e., GNT4, GMX6, IP3A, GMR10 and GTB17) were collected from a demo farm in Batac, Ilocos Norte, Philippines. The bulbs were made to grow roots which were subsequently used for DNA extraction and PCR analysis. Results showed that PCR amplified sequences of *rbcL* and *matK* genes were able to identify *A. sativum* varieties with ample discriminatory power to differentiate closely related species. The identified SRAP markers were also able to show a degree of polymorphism among the five local varieties. The *rbcL* sequences were highly conserved, the *matK*-sequences were able to discriminate GTB17 while the SRAP using the primer combinations ME1-EM3 and ME3-EM3 discriminated all the five varieties. A summary of the concatenated molecular features of the five varieties was provided. Albeit preliminary, the study was able to show the potential of *matK* and SRAP in discriminating the local *A. sativum* varieties from each other and from other varieties of different origin.

**Keywords:** *Allium sativum*, molecular characterization, *rbcL*, *matK*, SRAP

## 1. INTRODUCTION

*Allium sativum*, commonly known as garlic or “*bawang*” in the Philippines, is an herbaceous monocot plant most distinguishable by the bulb at its leaf base. It belongs to the family Amaryllidaceae and is believed to have originated from Central Asia, although it is now widely used worldwide as a food additive or spice (Quisimbing, 1978; Fritsch & Friesen, 2002; Block, 2010; Agarwal, 1996). It has also been known as a natural treatment for various illnesses or diseases as it exhibits antimicrobial, antiseptic, antitoxic, antiviral effects and acts as depurative, diuretic, expectorant, etc. (Agarwal, 1996,

Fufa, 2019). It contains chemically active compounds known for their therapeutic effects like flavonoids, saponins, essential oils, etc. and biologically active compounds which have antibiotic and fungicidal properties. In addition, it also has beneficial effects on the cardiovascular system and in cancer (Lanzotti, 2007; Banerjee & Maulik, 2002; Kumar *et al.*, 2010; Li *et al.*, 2018). Furthermore, *A. sativum* has also been found by some researches to have activities against insects, nematodes, rodents, and mollusks (Amonkar & Reeves, 1970; Singh & Singh, 2008; Nwanchukwu & Asawalam, 2014).

*A. sativum*, along with *A. cepa*, commonly known as onion, have varying characteristics such as size, color, flavor and more importantly, chemical components, making them important in traditional medicine, food additive, and in many other aspects (Boukeria *et al.*, 2016). The Philippine Bureau of Plant Industry identified six *A. sativum* varieties that are produced in the Philippines. These are the Ilocos White, Tan Bolters, Batanes White, Batangas White, Ilocos Pink and Nueva Ecija Pink. In addition to these, the Mariano Marcos State University in Batac, Ilocos Norte also produces the Native, Mexican and Miracle varieties. Different garlic varieties have different chemical profiles that may be responsible for pest and disease resistance, as well as its medicinal properties. Moreno *et al.* (2016) determined the difference in chemical profile and cytotoxic activity of two varieties, Ilocos White and Native garlic.

The Consortium of Barcode of Life (CBOL) has recommended *rbcL* for its universality, and *matK* for its discriminatory power (CBOL, 2009), and these have been used for the identification and phylogenetic analysis of *Allium* species (Ipek *et al.*, 2014; Lee *et al.*, 2017; Abugalieva *et al.*, 2017; Zarei *et al.*, 2020). Sequence-Related Amplified Polymorphism (SRAP) markers have also been demonstrated to reveal 69.1% polymorphic loci in 40 garlic germplasms from China (Chen, *et al.*, 2013). With these, the study aimed to evaluate the utility of chloroplast genes, *rbcL* and *matK*, as well as Sequence-Related Amplified Polymorphism (SRAP) to define and characterize five *A. sativum* varieties grown in the Philippines.

## 2. METHODOLOGY

### 2.1. Acquisition of Samples

Garlic varieties that were readily available at the time of sampling were collected from the demo farm of Mariano Marcos State University in Batac, Ilocos Norte, Philippines (Table 1). Bulb specimens for each variety were vacuum sealed until used for germination of the cloves. Further experimentation was conducted at the Department of Biochemistry and Molecular Biology (DBMB), University of the Philippines – Manila.

**Table 1.** Local garlic varieties collected from Batac, Ilocos Norte, Philippines that was used in this study

Sample Code	Identification	Locality	Genbank Accession No.
GNT4	Native		MN239994
GMX6	Mexican	Batac, Ilocos Norte, Philippines	MN239995
IP3A	Ilocos Pink		MN239996
GMR10	Miracle		MN239997
GTB17	Tan Bolters		MN239998

### 2.2 DNA Extraction

One garlic bulb specimen for each variety was used for DNA extraction. Garlic cloves were peeled and half-submerged in water using floaters to allow roots to grow.

About 1 cm of the root tips were cut and used for DNA extraction. DNA extraction was performed following the manufacturer's protocol for i-genomic Plant DNA Extraction Mini Kits (iNTRON Biotechnology, South Korea). Briefly, 50 mg of the root tips were homogenized in a 1.5 mL microcentrifuge tube using a micropestle. Binding, washing and elution of DNA was performed in a spin column. To verify the success of extraction, samples were subjected to 1.2% agarose gel electrophoresis with 0.5X TAE buffer and 1 µL GelRed and visualized through a gel documentation system (Bio-rad, USA). Garlic root DNA isolates were then stored in -20°C freezer until used for PCR amplification.

### 2.3 PCR Amplification, Agarose Gel Electrophoresis, DNA Sequencing, and Sequence-Related Amplified Polymorphism (SRAP) Analysis

Three markers (*rbcL*, *matK* and SRAP) were used in PCR amplification using the Bio-rad T-100 Thermocycler (USA). Genomic DNA was amplified using Promega (USA), and Invitrogen Platinum *Taq* Polymerase for *rbcL*/*matK* and SRAP, respectively. PCR reaction mixture and thermal cycle profile were set up according to the manufacturer's protocol. Primer sequences and thermal profile is shown in Table 2. The 30 µl PCR reaction mixture of the markers, *rbcL* and *matK* are as follows: 22.7 µL of nuclease-free water, 3 µL of 10X PCR buffer, 1.2 µL of 50 mM MgCl<sub>2</sub>, 0.6 µL of 10 mM deoxynucleotide triphosphates (dNTP), 0.6 µL of 10 mM forward and reverse primers, 0.3 µL of *Taq* DNA polymerase (5 U/ µL), and 1 µL of the DNA template (20 ng).

**Table 2.** List of primer pairs used to amplify the *rbcL*, *matK* and SRAP region. Primer combinations for SRAP were: ME1-EM3 and ME3-EM3; Thermal profile for each marker was also presented.

Marker	Primer name	Primer Sequence (5'-3')	Thermal Profile
<i>rbcL</i>	<i>rbcLa-F</i>	5'- ATGTCACCACAAAC AGAGACTAAAGC-3'	94°C for 4 minutes, 94°C for 30 seconds, 55°C for 30 seconds, 72°C for 1 minute, 35 cycles, and 72°C for 10 minutes.
	<i>rbcLa-R</i>	5'- GTAAAATCAAGTCC ACRCG-3'	
<i>matK</i>	<i>matK-xf</i>	5'- TAATTTACGATCAAT TCATTC-3'	94°C for 1 minute, 94°C for 30 seconds, 54°C for 20 seconds, 72°C for 50 seconds, 35 cycles and 72°C for 5 minutes.
	<i>matK-MALP</i>	5'- ACAAGAAAGTCGAA GTAT-3'	
SRAP	ME1	5'-TGA GTC CAA ACC GGA TA-3'	94°C for 5 minutes, 94°C for 1 minute, 35°C for 1 minute, and 72°C for 1 minute, 5 cycles; 94°C for 1 minute, 50°C for 1 minute, 72°C for 1 minute, 35 cycles, and 72°C for 5 minutes.
	ME3	5'-TGA GTC CAA ACC GGA AT-3'	
	EM3	5'-GAC TGC GTA CGA ATT GAC-3'	

Agarose gel electrophoresis was performed to verify the presence of amplified bands using 1.2% concentration with 0.5X TAE Buffer and 1  $\mu$ L of gel red, and run at 100V for 30 minutes. The gel containing the successfully amplified products was visualized using gel documentation system (Bio-Rad, USA) (Figures 1a and 2a). PCR amplicons were sent to Macrogen, South Korea for bidirectional sequencing.

SRAP was performed using twenty-five primer combinations to determine the polymorphic sites of each garlic variety. The 10  $\mu$ L PCR reaction mixture of the SRAP marker are as follows: 3.6  $\mu$ L of nuclease-free water, 5  $\mu$ L of Invitrogen Master Mix, 0.2  $\mu$ L of 10 mM forward and reverse primers, and 1  $\mu$ L of DNA template (20 ng). Agarose gel electrophoresis was performed to verify the presence of amplified polymorphic bands using 2.2% concentration with 0.5X TAE Buffer and 1  $\mu$ L of gel red and was subjected to electrophoresis at 75V for 45 minutes. Visualization was done using gel documentation system (Bio-Rad, USA).

#### 2.4 Sequence Analysis for *rbcL* and *matK*

Sequence assembly and alignment was done using Geneious Prime 2020 (<http://www.geneious.com/>) and MEGA X software (Kumar, Stecher, Li, Knyaz, and Tamura 2018), respectively. BLASTn analysis (Altschul, et al., 1990) was generated to determine the sequence homology to *Allium* sequences deposited at NCBI GenBank. Sequences generated were trimmed at both the 5' and 3' ends to obtain a consensus sequence with high quality base calls. *Allium schoenoprasum* (chives), *Allium ampeloprasum* (wild leek) and *Allium cepa* (onion) were used as outgroup taxa to evaluate the discriminatory power of these two markers. Maximum-Likelihood trees were generated from the resulting sequences using models determined by the *Find Best DNA/Protein Models (ML)* algorithm of MEGA X.

#### 2.5 SRAP Analysis

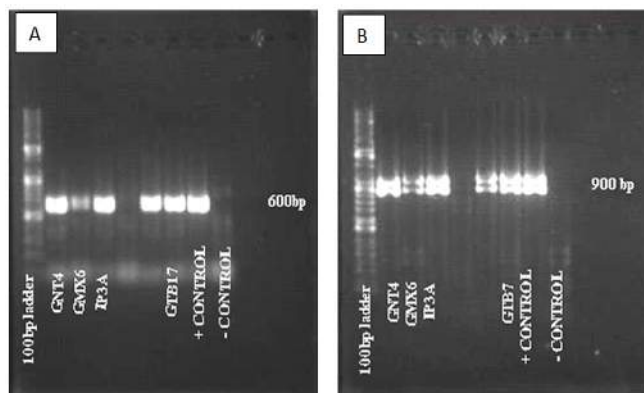
Presence (1) or absence (0) of distinct bands in the gel image were noted and compared among the samples used in this study as polymorphic markers (Li & Quiros, 2001).

### 3. RESULTS

#### 3.1 PCR Amplification of *rbcL* and *matK* genes

Amplicons of the *rbcL* gene were generated from the four local *A. sativum* varieties analyzed (Fig. 1). Approximately 600 bp amplicon was generated for GMX6, IP3A, GMR10 and GTB17. GNT4 sequence was obtained from a previous amplification experiment (Olivar *et al.*, unpublished). The five *rbcL* sequences were deposited in GenBank: GNT4 (MN239994), GMX6 (MN239995), IP3A (MN239996), GMR10 (MN239997) and GTB17 (MN239998) (Table 1).

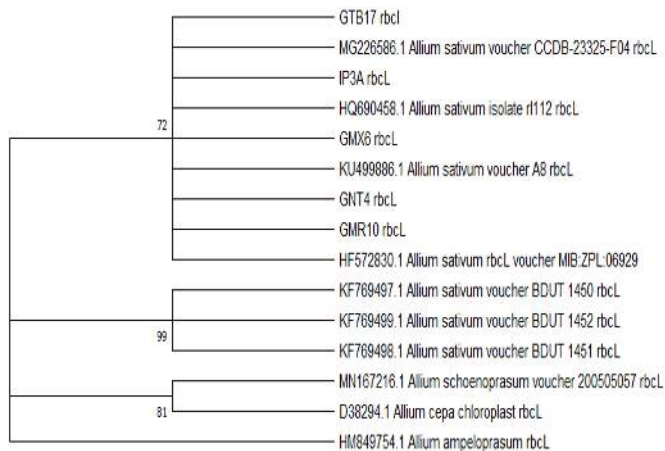
Sequences of the *rbcL* gene of the five local *A. sativum* varieties were compared. Upon alignment and trimming,



**Figure 1.** Gel profile showing PCR amplicons using (A) *rbcL* and (B) *matK* marker resolved in 1.2% agarose gel. A 100 bp Plus DNA ladder was used as molecular marker.

408 base pairs were obtained for length consistency in performing the phylogenetic analysis. The BLASTn analysis for the *rbcL* region generated a 99-100% identity match with *Allium sativum* sequences deposited at NCBI GenBank. A 99-100% sequence identity without other similarity to other species is an indicator of positive identification (Hofstetter et al., 2019).

Phylogenetic analysis was conducted on the five local *A. sativum* varieties used in this study, along with seven *A. sativum* reference sequences and three closely related *Allium* species as outgroup (Fig.2). Analysis was done through construction of the maximum-likelihood tree using the Jukes-Cantor model. All five *A. sativum* sequences examined in this study clustered to a single branch with the reference *A. sativum* sequences MG226586, HQ690458, KU499886 and HF572830.

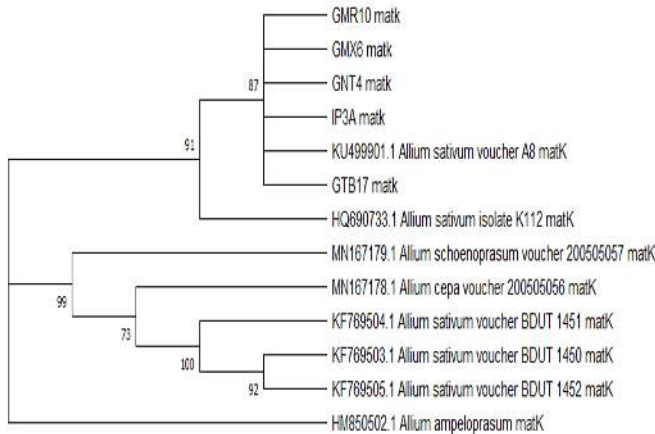


**Figure 2.** Maximum Likelihood tree of aligned *rbcL* sequences generated using the Jukes-Cantor model. The bootstrap consensus tree was inferred from 1000 replicates and condensed to >70% bootstrap values. There was a total of 408 positions in the final dataset. The tree was rooted to the outgroup *A. ampeloprasum*.

The *rbcL* region was not able to distinguish the five varieties from one another but were discriminated from three other *A. sativum* sequences KF769497,

KF769498 and KF769499, as well as from closely related *Allium* species *A. schoenoprasum*, *A. cepa* and *A. ampeloprasum*. A total of 54 variable sites were observed to be responsible for the branching of these three from the rest of the *A. sativum* sequences (Fig 5, A).

A similar approach was used for the analysis of the *matK* gene. Sequences were obtained from the *matK* region of five *A. sativum* variants (Figure 3). Upon alignment and trimming, 653 base pairs were obtained for length consistency in performing the phylogenetic analysis. The BLASTn analysis for *matK* region generated a 98.63-100% identity match with the *A. sativum* sequences deposited at NCBI GenBank.

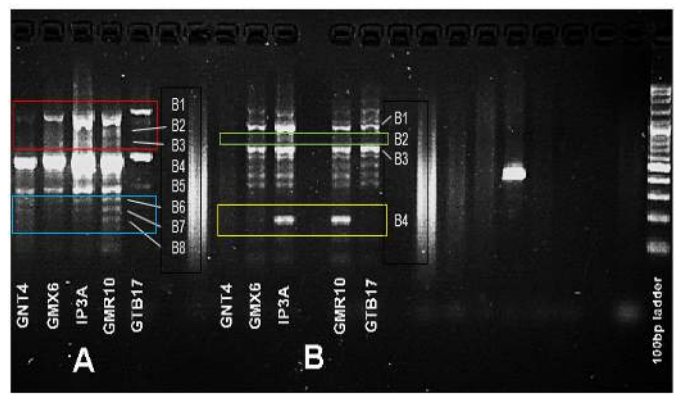


**Figure 3.** Maximum Likelihood tree of aligned *matK* sequences generated using the Tamura 3-Parameter model (T92). The bootstrap consensus tree was inferred from 1000 replicates and condensed to >70% bootstrap values. There was a total of 653 positions in the final dataset. The tree was rooted to the outgroup *A. ampeloprasum*.

Phylogenetic analysis was conducted on five local *A. sativum* varieties used in this study, along with five reference sequences and three closely related *Allium* species as outgroup. Analysis was done through construction of the maximum-likelihood tree using the Tamura-3-parameter model (T92). A total of 45 variable sites were observed in the *matK* sequence alignment of ten *A. sativum* species (Fig 5, B). All five *A. sativum* sequences examined in this study clustered with the reference sequence KU499901. Meanwhile, the other reference sequences, HQ690733, KF769503, KF769504 and KF769505 formed different branches. The outgroup sequences were also successfully discriminated from the *A. sativum* samples.

### 3.2 SRAP analysis

After visualization, out of twenty-five primer pairs only two yielded polymorphic bands, consequently, ME1-EM3 and ME3-EM3 primer pairs were chosen. Bands resolved after gel electrophoresis using SRAP markers were evaluated (Figure 4). Polymorphic bands at



**Figure 4.** Gel profile showing PCR amplicons from SRAP primers combinations (A) ME1-EM3 and (B) ME3-EM3 resolved in 2.2% agarose gel. Polymorphic bands at around 1200, 1000, 800 bp (red), 300,250,200 bp (blue), 900 (green), and 200 bp (yellow) were highlighted. A 100 bp Plus ladder was used as molecular marker.

around 1200, 1000, 800 bp (i.e., enclosed in red box), and 300,250,200 bp (i.e., enclosed in blue box) for ME1-EM3 were generated, and 900 (i.e., enclosed in green box), and 200 bp (i.e., enclosed in yellow box) for ME3-EM3 were generated. A distinct polymorphic band at around 200 bp was present in the IP3A and GMR10 varieties while a loss of the 1,200 bp band was noted in GNT4. IP3A and GMR10 had the greatest number of bands generated for both primers, GNT4 had the least, followed by GMX6 and GTB17. Interestingly, GNT4 did not produce bands for ME3-EM3. Either this was due to failed amplification or absence of the target polymorphic region, is subject to further investigation.

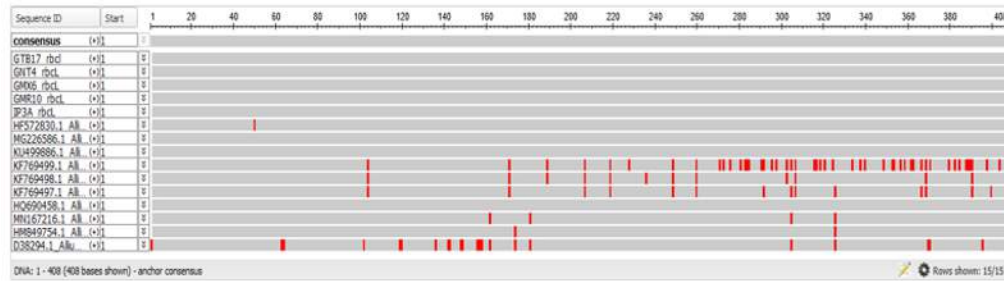
### 3.3 Sequence alignment and summary of concatenated molecular profiles

The sequence alignment for *rbcL* and *matK* sequences are summarized in Figure 5 together with the results of the SRAP. As can be noted, from the 408-nt *rbcL* sequences, no polymorphism is observed among the local varieties studied and are thus, highly conserved, although two of the closest relatives reported in Genbank showed differences in nucleotide numbers 50, 174, and 327, For the 653 nt *matK*-sequence, GTB17 was discriminated at 5 nucleotide positions (i.e., 4, 76, 85, 126 and 445). While the rest of the variants have similar *matK*-sequences, they differ from the closest relatives reported in Genbank in nucleotide numbers 245, 289, 509 and 535. Meanwhile, the SRAP discriminates all the five varieties based on the 8 bands generated from ME1-EM3 primers and 4 bands from ME3-EM3 primer pairs.

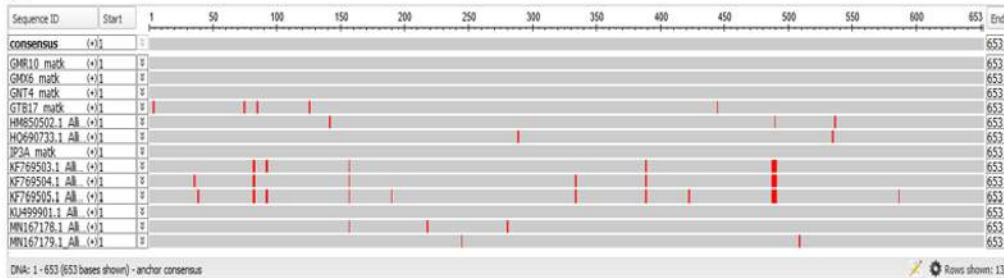
## 4. DISCUSSION

Species under the *Allium* genus can be distinguished from each other based on their morphological characteristics (Khosla, 2014). However, within species variation may prove to be difficult to distinguish based on morphological traits alone. Differentiation among *A. sativum* varieties is primarily based on average plant height, average bulb size and bulb color as described in

A.



B.



C.

Garlic varieties studied	rbcL sequence (1-408 nt)											matK sequence (1-653 nt)											SRAP							
	Nucleotide position number											Nucleotide position number											Band number for ME1-EM3				Band number for ME3-EM3			
	50	174	327	4	76	85	126	245	289	445	509	535	1	2	3	4	5	6	7	8	1	2	3	4						
GMR10	A	A	T	A	A	T	A	G	G	A	A	A	1	1	1	1	1	1	1	1	1	1	1	0	1	0				
GMX6	A	A	T	A	A	T	A	G	G	A	A	A	1	0	0	1	1	1	0	0	0	1	0	1	0	1	1			
GNT4	A	A	T	A	A	T	A	G	G	A	A	A	0	0	0	1	1	1	1	1	1	F	F	F	F	F				
IP3A	A	A	T	A	A	T	A	G	G	A	A	A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
GTB17	A	A	T	C	C	G	C	G	G	C	A	A	1	0	0	1	1	0	0	0	1	0	1	0	1	0	0			
HF572830.1 (GB)	C	A	T	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			
HM849754.1 (GB)	A	G	C	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			
HO690733.1 (GB)	N	N	N	A	A	T	A	G	A	A	A	G	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			
MN167179.1 (GB)	N	N	N	A	A	T	A	A	G	A	T	A	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			

**Figure 5.** Summary of the alignment profiles of the garlic varieties and their concatenated molecular features. A- *rbcL* alignment of 408 nt, B- *matK* alignment of 653 nt, and C- summary of concatenated molecular features. Legends: F-Failed amplification, N-No data, GB-Genbank. As highlighted in yellow in panel C, *rbcL* sequences are highly conserved, *matK*-sequences discriminates GTB17. SRAP discriminates all the five varieties.

URLA: [https://www.ncbi.nlm.nih.gov/projects/msviewer/?coloring=diff&consensus=true&key=kSIL-40gUvn-Dhz-3R8qAHqUrX0MbApDm8meTJ9lFOXyCo2nwnzWjP1By4hA\\_5XuED9R3RRtQBzhXDE8Uf3i33EPk8xRY\\_GKu C c g S p 2 3 B 3 h 5 V 3 V J a j i f M i 2 p K F 4 4 v m h - C v 9 r v y q d w 4 7 6 0 5 G Y M V Q J 2 l q K I A w 8 G - kKbNstOWiJPPiduE3YLRmeOw3r7ygtg&columns=d:120,b:50,x:17,aln.](https://www.ncbi.nlm.nih.gov/projects/msviewer/?coloring=diff&consensus=true&key=kSIL-40gUvn-Dhz-3R8qAHqUrX0MbApDm8meTJ9lFOXyCo2nwnzWjP1By4hA_5XuED9R3RRtQBzhXDE8Uf3i33EPk8xRY_GKu C c g S p 2 3 B 3 h 5 V 3 V J a j i f M i 2 p K F 4 4 v m h - C v 9 r v y q d w 4 7 6 0 5 G Y M V Q J 2 l q K I A w 8 G - kKbNstOWiJPPiduE3YLRmeOw3r7ygtg&columns=d:120,b:50,x:17,aln.)  
 URLB: [https://www.ncbi.nlm.nih.gov/projects/msviewer/?coloring=diff&consensus=true&key=ohE4yL4TYcrNPS\\_N7iwZM0mYU0AMMQl0DjlmJDIgIA6xPSpp9VM1rCP6FoRb90-KHpJdHl2iBqdBvVWwU7ZfrW2EUlp8tIY,VOfOPkjlzW7y9k7GNrvxb9umKTH1cnQxdbtwPnE6-p62eGNPbdH-2A3VSB3XxEiQDodLgMKWA8fQsYDR4BBTMsDilIHgg&columns=d:120,b:50,x:17,aln,e:50](https://www.ncbi.nlm.nih.gov/projects/msviewer/?coloring=diff&consensus=true&key=ohE4yL4TYcrNPS_N7iwZM0mYU0AMMQl0DjlmJDIgIA6xPSpp9VM1rCP6FoRb90-KHpJdHl2iBqdBvVWwU7ZfrW2EUlp8tIY,VOfOPkjlzW7y9k7GNrvxb9umKTH1cnQxdbtwPnE6-p62eGNPbdH-2A3VSB3XxEiQDodLgMKWA8fQsYDR4BBTMsDilIHgg&columns=d:120,b:50,x:17,aln,e:50)

the Philippine Department of Agriculture Garlic Production Guide (Department of Agriculture, 2012). Cultivars that are actually the same but with different names typically arise from this kind of phenotypic/agronomic traits classification (Egea et al., 2017).

In this study both the *rbcL* and *matK* DNA regions were shown to be highly conserved in *A. sativum*, making them suitable barcoding markers for the molecular identification and discrimination of *A. sativum* from other closely related *Allium* species.

Unlike *rbcL*, *matK* is highly variable and have been used to differentiate closely related species (Dong et al., 2012). The *matK* sequences obtained have correctly identified all the *A. sativum* samples at 98-100% match.

All five varieties showed sequence homology with the reference sequence KU499901 which presumably originated from Egypt based on its GenBank entry. However, the other, reference sequences HQ690733 (China), KF769503, KF769504 and KF769505 (India) showed clear distinction from the rest. This result suggests that the *matK* DNA marker exhibited both inter- and intra-species discriminatory power, albeit with some degree of resolution. Although unable to discriminate among varieties, this study has shown the existence of *A. sativum* species that have polymorphic *rbcL* and *matK* regions. In particular, garlic sequences deposited in GenBank that originated from China and India were distinct from the local varieties in the Philippines. Intraspecific variation in *matK* has also been reported in other species (Hori et al., 2006; Spies and Spies, 2018).

Sequence-Related Amplified Polymorphism (SRAP) markers are DNA-based dominant markers with primers designed to target open reading frames. (Li and Quiros, 2001). Two primer sets were evaluated to generate distinct banding patterns. Two polymorphic bands (900 bp and 200 bp) were observed from the ME3-EM3 primer set that was able to differentiate the varieties IP3 and GMR10 from the other three varieties. Meanwhile,

the ME1-EM3 primer was able to discriminate GNT4. This preliminary result on the use of SRAP markers show its potential to discriminate among varieties within *A. sativum* species, which may be used to augment the discriminatory power of *matK*.

## 5. CONCLUSION AND RECOMMENDATION

In this study, the utility of DNA markers to identify *A. sativum* species were shown. The *rbcL* and *matK* DNA regions was utilized as molecular markers for the species identification of *Allium sativum*. The *rbcL* sequences were highly conserved, the *matK*-sequences discriminated GTB17 while the SRAP using the primer combinations ME1-EM3 and ME3-EM3 discriminated all the five local garlic varieties. Albeit preliminary, the data suggest that the *matK* sequences and the SRAP showed potential as markers for augmenting the current phenotypic classification of *A. sativum*, and for discriminating varieties of the same *A. sativum* from different origins. An increase in sample size with more varieties are needed to validate these findings.

## ACKNOWLEDGEMENT

We would like to extend our gratitude to Mariano Marcos State University, Batac, Ilocos Norte for the *A. sativum* voucher specimens and the Department of Biochemistry and Molecular Biology (DBMB), University of the Philippines–Manila, for the use of the laboratory facilities.

## REFERENCES

1. Abugalieva, S., Volkova, L., Genievskaya, Y., Ivaschenko, A., Kotukhov, Y., Sakauova, G., & Turuspekov, Y. (2017). Taxonomic assessment of *Allium* species from Kazakhstan based on ITS and *matK* markers. *BMC Plant Biology*, 17(S2). doi:10.1186/s12870-017-1194-0
2. Agarwal, K. C. (1996). Therapeutic actions of garlic constituents. *Medicinal Research Reviews*, 16(1), 111-124. doi:10.1002/(sici)1098-1128(199601)16:13.0.co;2-5
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
4. Amonkar, S. V., & Reeves, E. L. (1970). Mosquito Control with Active Principle of Garlic, *Allium sativum*. *Journal of Economic Entomology*, 63(4), 1172-1175. doi:10.1093/jee/63.4.1172
5. Banerjee, S. K., & Maulik, S. K. (2002). Effect of garlic on cardiovascular disorders: A review. *Nutrition Journal*, 1(1). doi:10.1186/1475-2891-1-4
6. CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31), 12794–12797. <https://doi.org/10.1073/pnas.0905845106>
7. Chen, S., Zhou, J., Chen, Q., Chang, Y., Du, J., & Meng, H. (2013). Analysis of the genetic diversity of garlic (*Allium sativum* L.) germplasm by SRAP. *Biochemical Systematics and Ecology*, 50, 139-146. doi:10.1016/j.bse.2013.03.004
8. Dong, W., Liu, J., Yu, J., Wang, L., & Zhou, S. (2012). Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. *PLoS ONE*, 7(4). doi:10.1371/journal.pone.0035071
9. Egea, L. A., Mérida-García, R., Kilian, A., Hernandez, P., & Dorado, G. (2017). Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology “Genotyping-by-Sequencing” Platform (DArTseq). *Frontiers in Genetics*, 8. doi:10.3389/fgene.2017.00098
10. Fritsch, R. M., & Friesen, N. (n.d.). Evolution, domestication and taxonomy. *Allium Crop Science: Recent Advances*, 5-30. doi:10.1079/9780851995106.0005
11. Fufa, B. K. (2019). Anti-bacterial and Anti-fungal Properties of Garlic Extract (*Allium sativum*): A Review. *Microbiology Research Journal International*, 1-5. doi:10.9734/mrji/2019/v28i330133
12. Hofstetter, V., Buyck, B., Eyssartier, G., Schnee, S., & Gindro, K. (2019). The unbearable lightness of sequenced-based identification. *Fungal Diversity*, 96 (1), 243-284. doi:10.1007/s13225-019-00428-3
13. Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and Using a Plant DNA Barcode. *PLoS ONE*, 6(5). doi:10.1371/journal.pone.0019254
14. Hori, T., Hayashi, A., Sasanuma, T., & Kurita, S. (2006). Genetic variations in the chloroplast genome and phylogenetic clustering of *Lycoris* species. *Genes & Genetic Systems*, 81(4), 243-253. doi:10.1266/ggs.81.243
15. Ipek, M., Ipek, A., & Simon, P. W. (2014). Testing the utility of *matK* and ITS DNA regions for discrimination of *Allium* species. *Turkish Journal Of Botany*, 38, 203-212. doi:10.3906/bot-1308-46
16. Khosa, J. S., Dhatt, A. S., & Negi, K. S. (2014). Morphological Characterization of *Allium* spp. Using Multivariate Analysis. *Morphological Characterization of Allium Spp. Using Multivariate Analysis*, 27(1), 24-27.
17. Kumar, K. P., Bhowmik, D., Chiranjib, & Biswajit. (2010). Aloe vera: A potential herb and its medicinal importance. *Journal of Chemical and Pharmaceutical Research*, 2(1), 21-29.
18. Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, 35(6), 1547-1549. doi:10.1093/molbev/msy096
19. Lanzotti, V. (2006). The analysis of onion and garlic. *Journal of Chromatography A*, 1112(1-2), 3-22. doi:10.1016/j.chroma.2005.12.016



20. Lee, J., Chon, J., Lim, J., Kim, E., & Nah, G. (2017). Characterization of Complete Chloroplast Genome of *Allium victorialis* and Its Application for Barcode Markers. *Plant Breeding and Biotechnology*, 5(3), 221-227. doi:10.9787/pbb.2017.5.3.221
21. Li, G., & Quiros, C. F. (2001). Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: Its application to mapping and gene tagging in Brassica. *Theoretical and Applied Genetics*, 103(2-3), 455-461. doi:10.1007/s001220100570
22. Li, Z., Le, W., & Cui, Z. (2018). A novel therapeutic anticancer property of raw garlic extract via injection but not ingestion. *Cell Death Discovery*, 4(1). doi:10.1038/s41420-018-0122-x
23. Moreno, P. G., Wang, L. Y., Alvarez, M. R., Olivar, D. D., Echavez, M. J., Yu, G. B., . . . Heralde, F. M. (2016). Preliminary Comparative Chemical Profiles and Cytotoxic Activities of Two Philippine *Allium sativum* Linn. Varieties: Ilocos White and Native. *Philippine Journal of Health Research and Development*, 20(2).
24. Nwachukwu, I. D., & Asawalam, E. F. (2014). Evaluation of freshly prepared juice from garlic (*Allium sativum* L.) as a biopesticide against the maize weevil, *Sitophilus zeamais* (Motsch.) (Coleoptera: Curculionidae). *Journal of Plant Protection Research*, 54(2), 132-138. doi:10.2478/jppr-2014-0021
25. Singh, D. K., & Singh, V. K. (2008). Pharmacological effects of garlic (*Allium sativum* L.). *Annual Review of Biomedical Sciences*, 10(0). doi:10.5016/1806-8774.2008.v10p6
26. Spies, J. J., & Spies, P. (2018). Assessing *Clivia* taxonomy using the core DNA barcode regions, matK and rbcLa. *Bothalia*, 48(1). doi:10.4102/abc.v48i1.2025
27. Yasmin, H., Anbumalarmathi, J., & Sharmili, S. A. (2018). Phytochemical analysis and antimicrobial activity of garlic (*Allium sativum* L.) and onion (*Allium cepa* L.). *Research on Crops*, 19(2), 245. doi:10.5958/2348-7542.2018.00035.9
28. Zarei, H., Fakheri, B. A., Naghavi, M. R., & Mahdinezhad, N. (2020). Phylogenetic relationships of Iranian *Allium* species using the matK (cpDNA gene) region. *Journal of Plant Biotechnology*, 47(1), 15-25. doi:10.5010/jpb.2020.47.1.015

# Preliminary characterization and *in silico* studies on the alpha-amylase inhibitors from *Momordica charantia* AMP06 methanolic leaf extract

Arra B. Asejo<sup>1\*</sup>, Patrick G. Moreno<sup>1\*</sup>, Junie Billones<sup>2</sup>, Ruel Nacario<sup>3</sup>, Francisco M. Heralde III<sup>1</sup>

<sup>1</sup>Department of Biochemistry and Molecular Biology, University of the Philippines Manila, Pedro Gil St., Ermita, Manila; <sup>2</sup>Department of Physical Science and Mathematics, University of the Philippines Manila, Padre Faura St., Ermita, Manila; <sup>3</sup>Department of Chemistry, University of the Philippines Los Baños, Laguna

## Email address:

\*arra.asejo@gmail.com; moreno.patrickgabriel@gmail.com

## To cite:

Asejo AB, Moreno PG, Billones J, Nacario R and Heralde FIIM.2020. Preliminary characterization and in silico studies on the alpha-amylase inhibitors from *Momordica charantia* AMP-06 methanolic leaf extract. PJBMB. Vol. I, No. 1, 2020, pp. 61-70. doi: 10.5555/pjbmb.ph.2020.01.01.61

Received: 07 25, 2019; Accepted: 12 18, 2020; Published: 01 25, 2021

**Abstract:** Regulating the activity of alpha-amylase, can help maintain blood glucose to normal levels especially among diabetics. *Momordica charantia*, known as ampalaya in the Philippines, was previously shown to have alpha-amylase inhibitory activity in our laboratory where among the nine locally available varieties, AMP06 gave the highest activity. This study aimed to characterize the alpha- amylase inhibitory principle in the leaf extract of AMP06 and perform *in silico* docking studies on putatively identified compounds. Crude methanolic leaf extract (CMLE) of *M. charantia* (AMP06) was prepared, subjected to sequential solvent extraction, concentrated by rotary evaporation, and tested for alpha-amylase inhibition. CMLE was shown to have inhibitory activity against alpha-amylase with values ranging from 43.53% - 65.70% at 500 ug/ mL to 25 ug/mL respectively. The sequential solvent partitioning of CMLE yielded fractions (i.e., hexane (HF), chloroform (CF), ethyl acetate (EAF) and butanol (BF) with decreased inhibitory activities that ranged from 20.12% to 38.02% compared with the CMLE. Fraction mixture of 1:1 ratio was prepared and evaluated and showed a recovery of inhibitory activity of 29.88% to 54.98%, indicating a possible synergistic activity. The individual fractions from the best fraction mixture were analyzed using LCMS/MS to determine the putative identity of the compounds which were then subjected to molecular docking to predict possible synergistic activity. Utilizing acarbose as a point of reference, as well as the top five hits from the docking experiment, *in silico* docking done with the compounds in the active site alone and in three allosteric sites, showed stronger binding affinity values in the active site of the enzyme. In conclusion, this study characterized the alpha-amylase inhibitory principle of *M. charantia* and its possible mechanism, that there may be compounds present in the extract that works synergistically in inhibiting alpha-amylase.

**Keywords:** Alpha-amylase, *Momordica charantia*, Ampalaya, synergistic activity

## 1. INTRODUCTION

Several strategies have been utilized to manage diabetes including diet (i.e., control of sugar intake), exercise, insulin injection and blood sugar lowering medication and herbal supplements (American Diabetes Association, 2017). One way to control post-prandial spike of blood glucose in people with diabetes is through the inhibition of alpha-amylase (Agarwal and Gupta, 2016), an enzyme released by the pancreas to

breakdown starch into limit dextrins that will be eventually cleaved to glucose for intestinal absorption. Alpha-amylase inhibitory activities have been reported in different plant extracts. *Mallotus repandus* (Willd.) stem extract showed a positive inhibitory activity against alpha-amylase, with the activity associated with various polyphenols (Hasan *et. al.*, 2014). Aqueous leaf and stem extract of *Coccinia grandis* inhibits alpha-amylase possibly due to the presence of flavonoids and phenols (Pultubr *et. al.*, 2017). Extract of *Atrocarpus*

*heterophyllus* showed high inhibitory activity to alpha-glucosidase and alpha-amylase, possibly due to the presence of flavonols or phenolic acids (Nair, Kavrekar and Mishra, 2013). *Phaseolus vulgaris* has a lot of varieties, but the variety with the potent alpha-amylase inhibitor was the white kidney bean (Tadros, Eryan, Yussef and Taha, 2014), and therefore was used in this study as a comparator. Inhibitory activity of this species was discovered in 1945 by Bowman. There are three isoforms of the amylase inhibitor from this variety, the  $\alpha$ AI-1,  $\alpha$ AI-2 and an inhibitor-like isoform,  $\alpha$ AIL (Obiro, Zhang and Jiang, 2008). A commercial product based on this white kidney bean extract (WKBE) has been shown to reduce carbohydrate absorption by as much as 66% and regarded as generally recognized as safe (GRAS) food supplement by the US FDA (<https://starnpage.com/app/global/campaign/sampling-detail/209>).

*Momordica charantia*, locally known as ampalaya, is a common kitchen vegetable that has also been subject to various in vitro studies because of its folkloric use as a remedy for diabetes (Anilakumar, Kumar and Ilaiyara, 2015). In the study of Ee Shian *et. al.* (2015) and Guir (2016), *M. charantia* fruit extracts have shown inhibitory activity against alpha-amylase and alpha-glucosidase which was said to be comparable to acarbose. Gines *et. al.* (2013) characterized the inhibitory activity of the crude leaf extract of *M. charantia* against porcine pancreatic amylase by determining its effects on its  $K_m$  and  $V_{max}$ . Their results have shown possible uncompetitive inhibitory activity due to the decreased  $K_m$  and  $V_{max}$ . The study also suggested that the inhibitory activity of *M. charantia* must be further characterized.

Our laboratory investigated the alpha-amylase inhibitory activity of the nine varieties of local ampalaya. Among these varieties, AMP-06 showed the highest alpha-amylase inhibitory activity (Heralde *et. al.* 2017). This study aimed to characterize the alpha-amylase inhibitory principle of *M. charantia* AMP-06 and perform in silico docking studies on the putatively identified compounds to screen possibly synergistically acting combinations.

## 2. METHODOLOGY

### 2.1. Plant Material Preparation

The dried plant leaves of *M. charantia* variety AMP-06 (McAMP-06) from the PhilMetab Project (Heralde *et. al.*, 2017) was cut into small pieces and was subjected to grinding to a greater surface area available for the extraction method.

### 2.2 Extraction

The McAMP-06 leaves were submerged in 100% Methanol. Soaking was done for 16-18 hours in a stoppered Erlenmeyer flask at room temperature. After soaking, the extract will be filtered. The soaking procedure was done three times to maximize the

metabolite extraction. The filtrate was concentrated using a rotary evaporator in room temperature. To ensure dryness, the extract was placed in a dry bath. The extract was stored in a tightly sealed container in -20 degrees Celsius until used (Heralde *et. al.*, 2017).

### 2.3 Sequential Solvent Partitioning Method

The crude extract was subjected to sequential solvent partitioning using four solvents of different polarities – hexane, chloroform, ethyl acetate, and butanol. The dried methanolic leaf extract was reconstituted using 400 mL 95% methanol and placed in a separatory funnel. Partitioning was done twice for every solvent. The four fractions collected were concentrated using a rotary evaporator and placed in a dry bath to ensure dryness. The dried fractions were stored in -20°C until used (Riaz *et. al.*, 2012).

### 2.4 Alpha-amylase Inhibition Assay

Prior to the assay, the fractions were reconstituted using sodium phosphate buffer solution, pH 6.8, having the concentrations of 500  $\mu$ g/mL, 250  $\mu$ g/mL, 100  $\mu$ g/mL, 50  $\mu$ g/mL, and 25  $\mu$ g/mL. The extracts were subjected to vortex mixer to ensure homogeneity. The test was done in three trials with three replicates each.

The assay will be done in a 3-mL test tubes or 1.5-mL microcentrifuge tubes. For the test tubes containing the extract or fractions, 50  $\mu$ L phosphate buffer, 10  $\mu$ L porcine pancreatic alpha-amylase solution (2U/mL) and 20  $\mu$ L of the fraction was incubated at 37°C for 20 minutes. After incubation, 20  $\mu$ L of starch was added and incubated again for 30 minutes. 100  $\mu$ L of the color reagent (dinitro salicylic acid, DNS) was added and was placed in a water bath with a temperature of 100°C for 10 minutes. The tube was allowed to cool before transferring the contents in a 96-well plate for reading the absorbance at 540 nm in a Spectrostar Nano microplate reader (Heralde *et. al.*, 2017).

As for the negative control, 70  $\mu$ L of the phosphate buffer was placed on the tubes, no extract, fraction, acarbose or white kidney bean extract was added.

The positive control was acarbose with the same concentrations of the test extract or fraction (500  $\mu$ g/mL, 250  $\mu$ g/mL, 100  $\mu$ g/mL, 50  $\mu$ g/mL, and 25  $\mu$ g/mL).

Blank sample tubes were also prepared. The blank samples do not contain the enzyme of interest.

The results is presented by their percent inhibition using the following formula

$$I\alpha = \frac{Anc - A_{sample}}{Anc} \times 100$$

Where:  $I\alpha$  = percent inhibition  
Anc = net absorbance of the negative control

To calculate the net absorbance of each sample, the following equation was used:

$$A_{sample} = A_{test} - A_{blank}$$

Where: A<sub>sample</sub> = net absorbance of the test well  
A<sub>test</sub> = absorbance of the test well  
A<sub>blank</sub> = absorbance of the blank sample

### 2.5 LC-MS/MS Compound Identification

The volume of the purified fraction was adjusted to 10 mL with LCMS grade methanol. The fraction was filtered using a 0.2 µm PTFE syringe filter before injection. The analysis was carried out using Waters UPLC I-Class coupled with Xevo G2-XS Qtof Mass spectrometer (Waters Corp., USA). The separation of the compounds present in the fraction was performed by using a 2 x 100 mm 1.8-Micron Acquity HSS T3 column. Formic acid in water (0.1%) was the mobile phase A and the mobile phase B was 0.1% of formic acid in acetonitrile. Linear gradient of 5% mobile phase B to 95% was used for 15 minutes. The injection volume was 1 µL and the flow rate will have been 0.5 mL/min. The parameters that were used are the following: 2KV capillary voltage, 120°C source temperature, 500°C desolvation temperature, 40V cone voltage, 50 L/hr cone gas flow, 950 L/hr desolvation gas flow. Mass spectrometry data was acquired using positive ESI with m/z range of 50-1, 200 and was processed through UNIFI Scientific Information System v. 1.8 (Waters Corp., USA).

The resulting chromatogram compared with chromatogram of the solvents to remove the same peak and was then submitted online on different databases to find a matching compound and generate its structure. Compounds that were a good match, has the lowest mass error in terms of mDa and ppm, and has an available structure online were used for the docking studies. These compounds were presented using a 4 digit code of their PubChem ID.

### 2.6 Molecular docking studies

#### 2.6.1 Software

The molecular docking studies have been performed using PyRx v. 0.9.8 in an iMac 14.3 using a processor of 3.1 GHz 4-core Intel® Core™ i7 and memory of 8GB of 1600MHz DDR3. Discovery studio 2017 was used to see the interaction diagrams of the ligands and the protein in the specific binding sites.

#### 2.6.2 Structure files preparation

The 3D structure of porcine pancreatic alpha-amylase (PPA) was downloaded from the Research Collaboration for Structural Bioinformatics Protein Data Bank (RCSB PDB; www.rcsb.org). The PDB code for the crystal structure of PPA is 1OSE. After the preparation using PyMOL, the protein and the ligands were minimized to obtain their lowest energy conformations.

The 3D structure of the compounds present in the best fraction mixture was downloaded from PubChem.

PyMOL v. 2.3.1 was used for the preparation of the enzyme for docking and to measure the root-mean-square-deviation (RMSD) values. Water molecules and ligands of 1OSE were removed. The known inhibitor of

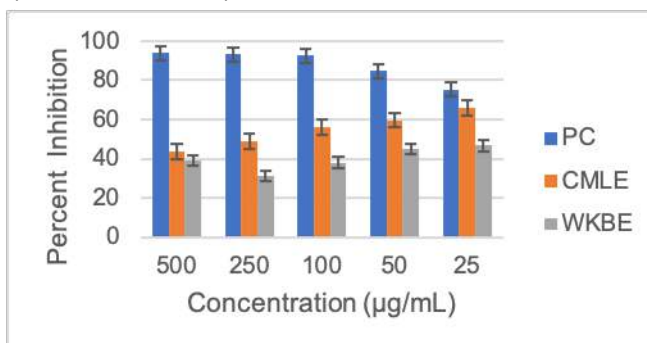
1OSE was redocked onto its binding site using appropriate Vina gridbox. The RMSD of the redocked and original inhibitor was calculated to validate the binding affinity of the standard.

### 2.7 Statistical Analysis

One-way analysis of variance (ANOVA) was used to determine the difference between the test groups and Tukey's Post-Hoc test was used to determine the concentration the highest activities. Dunnett's Test was utilized to compare the inhibitory activities of the fractions with the positive controls which is acarbose. Statistical results were considered to have a significant difference if the p-value ≤ 0.05, otherwise considered to be non-significant.

## 3. RESULTS

The crude methanolic leaf extract of McAMP-06 (CMLE) showed inhibition of alpha amylase (p-value = 0.009) with inhibitory activities ranging from 43.53% to 65.70%, compared with the inhibitory activity of the white kidney bean extract (WKBE) that ranges from 35.11% to 43.69% (Figure 1). Tukey's pos-hoc test revealed that there is no significant difference between the inhibitory activities between the concentrations of the crude extract. The Dunnett's post-hoc test showed that there is significant difference between the inhibitory activity of CMLE compared with the positive control which is acarbose (75.41% - 93.77%).



**Figure 1.** Alpha amylase inhibitory activity of crude methanolic leaf extract of *M. charantia* AMP-06 (CMLE) against white kidney bean extract (WKBE) and positive control, acarbose.

\* significant difference between acarbose and CMLE

\*\* significant difference between CMLE and WKBE

CMLE was subjected to fractionation through sequential solvent partitioning using four solvents with increasing polarities, hexane, chloroform, ethyl acetate and butanol. Three concentrations were made as these concentrations exhibited inhibitory activities higher than 50%.

It was observed that the fractions decreased in their inhibitory activity compared with CMLE (Figure 2). The inhibitory concentrations of hexane (HF), chloroform (CF) and ethyl acetate (EAF) fractions showed no significant difference against each other and against WKBE, while compared with butanol fraction (BF), it



acetate fraction and 14 are unique in the butanol fraction.

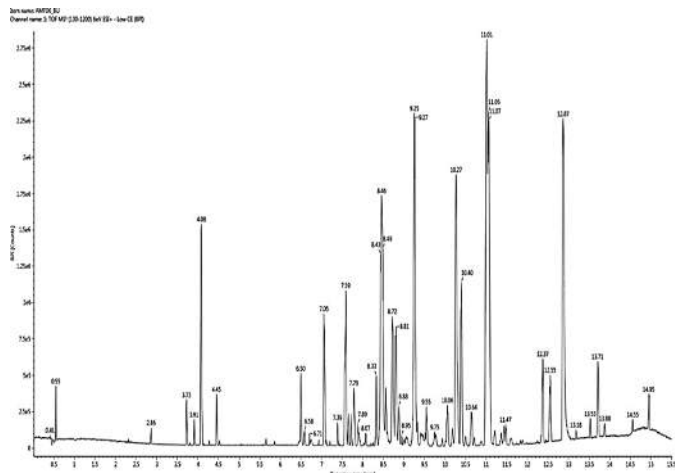


Figure 5. Base peak chromatogram of the Butanol fraction.

A ligand-enzyme docking approach involving three allosteric sites with the parameter settings outlined in Table S1 (Supplemental Table and Figures) was used. Two levels of evaluation were done, one involving single docking of the ligand with acarbose in the active site, and the other involving multiple binding in the three allosteric sites.

Table 1 shows the top five results of the docking of the ligands and acarbose alone in the active site. The ligands with high hits are Compounds 1889, 6324, 6169, 8554 and 0179. Four of these compounds are common in both fractions (i.e., a, b, c, and d) and one is unique to butanol fraction (i.e., 0179, e) (Table 2, underlined).

#### 4. DISCUSSION

This study was able to characterize the inhibitory activity of the AMP-06 variety of *M. charantia*. The crude extract produced an inhibitory activity ranging from 43.53% to 65.70%. This inhibitory activity of *M. charantia* crude extracts has been reported in many studies already. Beidokhti and Jager (2017) described that *M. charantia* has shown an inhibitory activity against alpha-amylase and alpha-glucosidase *in vitro* and exerted a hypoglycemic effect *in vivo*. Wang *et. al* (2019) described that *M. charantia* has been used as an anti-diabetic supplement for decades and in their study, they have evaluated the alpha-amylase and alpha-glucosidase inhibitory activity of freeze-dried and fresh *M. charantia*. They showed no significant difference on whether how *M. charantia* was prepared, both procedures resulted to inhibition of the enzymes. Wang *et. al* (2019) has suggested that this inhibitory activity was due to the presence of phytochemicals such as triterpenes and phenolic compounds.

**Table 1.** Top five compounds present in the ethyl acetate and butanol fractions with their structures and binding affinities to alpha amylase relative to acarbose.

Ligand	Structure	Binding affinity
1889		-9.5
0179		-9.2
6169		-8.9
8554		-8.6
6324		-8.3
Acarbose		-9.9

In so far as the specific compounds responsible for the alpha-amylase inhibition has not yet identified, Mc06-CLE was subjected to sequential solvent partitioning using hexane, ethyl acetate, chloroform, and butanol.

Table 2. List of the putative compounds with ID numbers in the ethyl acetate and the butanol fraction.

Ethyl acetate fraction		Butanol fraction	
Similar Compounds			
<u>6079</u>	3357	<u>6079</u>	3357
<u>6169<sup>c</sup></u>	<u>8554<sup>d</sup></u>	<u>6169<sup>c</sup></u>	<u>8554<sup>d</sup></u>
<u>6324<sup>b</sup></u>	7827	<u>6324<sup>b</sup></u>	7827
9215	2285	9215	2285
7911	4901	7911	4901
<u>2341</u>	1649	<u>2341</u>	1649
9251	<u>6433</u>	9251	<u>6433</u>
8544	2549	8544	2549
2591	<u>1889<sup>a</sup></u>	2591	<u>1889<sup>a</sup></u>
Unique Compounds			
<u>2339</u>		8545	<u>6269</u>
3788		0303	1828
		985	7037
		6009	0180
		<u>0179<sup>e</sup></u>	<u>2560</u>
		8553	<u>8241</u>
		5786	6272

Note: Underlined compounds, a-e bind to the active site. Compounds with similar colors and side colors were simultaneously bound to the allosteric sites and showed highly negative binding affinities (i.e., strong binding that could disrupt binding of the substrate and the positive control inhibitor, acarbose).

When the different components of the extract were separated based on their polarity, their inhibitory activities decreased. One mechanism that may explain how the inhibitory activity decreases once fractionated could be due to phytochemicals that protect the active enzyme from degradation (Efferth and Koch, 2011). This led to the hypothesis that these fractions may have synergistic effect with one another. This hypothesis was verified when the fractions were mixed with another fraction, where the inhibitory activity increases, from a range of 20% to 38% to a range of 32% to 54%. This result showed parallelism with the findings of He *et al.* (2016). Their study investigated the antioxidant activity, of individual and mixture, of three compounds present in *Cornus officinalis*. These compounds were loganin, morroniside and ursolic acid. Individually, morroniside, loganin and ursolic acid has increased the superoxide dismutase activity, an enzyme considered as a maker of oxidative stress (109.1 NU/mL, 110.6 NU/mL and 135.2 NU/mL respectively). When the ursolic acid and loganin was combined, SOD activity increased to 138.2 NU/mL.

Another study that shows synergism is by Liu *et al.* (2016) that examined the inhibitory activity of fractions from the water extract of Qingzhan dark tea. The crude water extract exhibited an IC<sub>50</sub> of 2.47 mg/mL, and after fractionation using three solvents, (chloroform, ethyl acetate and butanol) the chloroform fraction had no

inhibitory activity against alpha-glucosidase while ethyl acetate and butanol fractions had IC<sub>50</sub> of 2.27 mg/mL and 2.94 mg/mL respectively. This shows that ethyl acetate had the highest inhibitory activity among the fractions, but the correlation analysis between the IC<sub>50</sub> values and the total theaflavins and polyphenols, showed that the inhibitory activity was likely to be the result of the synergistic interaction of the phytochemicals mentioned in the said plant.

In this study, the data suggest a different direction about the conventional drug development process, that is a reductionistic method, where selective ligands are designed to act on a single disease target. Apparently, plant extracts must be studied not solely for having to isolate single compounds but to analyze their pharmacological effect based on their synergistic or antagonistic interactions (Efferth and Koch, 2011).

In silico studies has been done to see the interaction of the compounds with the allosteric and active site of alpha-amylase. As shown in the results, the compounds present in the ethyl acetate-butanol mixture work with each other to produce a more potent inhibition compared with the inhibition produced by having a single compound in the active site.

Figure S1 (Supplemental Table and Figures) shows the interaction of the known inhibitor of alpha-amylase in its active site. This inhibitor was redocked to validate the docking protocol used. The RMSD value between the original and the redocked ligand was 0.699, which is an acceptable value, because it is below 1.5. The main amino acids responsible for breaking down starch into limit dextrins are Asp197, Glu233 and Asp300. The figure shows that acarbose interacts with Asp300, Asp197, Glu233 and <sup>others</sup>, by a conventional hydrogen bond, and having a carbon hydrogen bond with Tyr62. Van der Waals interaction is found in many residues having a binding affinity of -9.9. This shows that acarbose competitively inhibit by interacting with the three amino acid residues that breaks starch into smaller carbohydrate molecules.

For compound 1889 in the active site, interactions with Asp197, Glu<sup>233</sup> and Asp<sup>300</sup>, are notable which could be responsible for the enzyme's mechanism via conventional hydrogen bond. For the compounds shown in Figures S2-S6, they only interact with Asp197, Glu<sup>233</sup> and Asp<sup>300</sup> through Van der Waals interaction.

There is a total of seven combination of compounds present in the ethyl acetate-butanol mixture that illustrate how this mixture inhibits alpha- amylase. As phytochemicals bind to the allosteric sites, the binding affinity of 2341 decreased to -10.6, comparable to the binding affinity of 1889 and acarbose that is -9.5 and -9.9 respectively. This decrease in binding affinity could be due to the interaction of 2341 to many amino acid residues in the active site, as shown in Figure S7, that includes Asp197, Glu233, Asp300 and Trp357, to which

Van der Waals and conventional hydrogen bonds, respectively are involved. Figures S8 and S9 are the two other combination of compounds that produce more negative binding energy than acarbose.

For the top hit found in Figure S7, 8241 is found in the butanol fraction only and while 2341 is found in the ethyl acetate fraction only, and 2339 is found in both fractions. Having these three compounds actively bind to the allosteric sites and another compound in the active site may afford the synergistic inhibition of alpha-amylase.

## 5. CONCLUSION AND RECOMMENDATION

This study characterized the inhibitory activity of the crude methanolic leaf extract of *M. charantia* AMP-06 (CMLE) variety against alpha- amylase, an enzyme responsible in breaking down starch into limit dextrins. The activity of this enzyme contributes to the postprandial hyperglycemia in diabetic patients.

The inhibitory activity of the CMLE, its fractions and the fraction mixture were determined by an alpha-amylase inhibition assay. Fractionation was done by sequential solvent partitioning using four solvent, hexane, chloroform, ethyl acetate, and butanol. Since there was a decrease in the inhibitory activity after fractionation, six fraction mixtures were made with 1:1 ratio. Fraction mixtures had an increase in their inhibitory activities, implying that fractions work together to inhibit the enzyme of interest.

The ethyl acetate-butanol mixture had the highest inhibitory activity among the mixtures and was subjected to LC-MS/MS analysis to determine the compounds present in the mixture. The compounds with known structure were docked using PyRx software to show how the compounds present in the mixture work together. In silico docking of individual phytochemicals, acarbose and in combination in three allosteric sites, was able to shortlist candidate phytochemicals that may work together synergistically in inhibiting alpha amylase. Further studies would require isolation of the pure compounds and validating the in-silico results in in-vitro enzyme inhibition assays. Also, the kinetics for this kind of synergistic combinatorial interaction need to be developed to be able to understand the mechanism for this type of inhibition. These compounds may also be evaluated for their potential utility as metabolite biomarker in identifying varieties of *M. charantia* with potent alpha-amylase inhibiting activity.

## REFERENCES

1. Agarwal, P., & Gupta, R. (2016). Agarwal, P. and Gupta, R. 2016. Alpha-amylase inhibition can treat diabetes mellitus. *Research and Reviews Journal of Medical and Health Sciences*, 5(4), 1-8.
2. American Diabetes Association. Standards of Medical Care in Diabetes – 2017. *Diabetes Care*. 40 (1): 51 – S135.
3. Anilakumar, K. R. (2015). Nutritional, Pharmacological and Medicinal Properties of *Momordica charantia*. *International Journal of Nutrition and Food Sciences*, 4(1), 75. doi:10.11648/j.ijnfs.20150401.21
4. Beidokhti, M.N. and Jager, A.K. 2017. Review of antidiabetic fruits, vegetables, beverages, oils and spices commonly consumed in the diet. *Journal of Ethnopharmacology*. 201: 26-41.
5. Ee Shian, T., Kassim, K., & Za, S. (2015). Antioxidant and hypoglycemic effects of local bitter gourd fruit (*Momordica charantia*). *International Journal of PharmTech Research*, 8(1), 974-4304.
6. Efferth, T. and Koch, E. 2011. Complex interactions between phytochemicals. The multi-target therapeutic concept of phytotherapy. *Current Drug Targets* 12 (1): 122-132.
7. Gines, K. R., Lee, A. and Lazaro-Llanos, N. 2013. The Effects of *Momordica charantia* Crude Leaf Extract on the Enzyme Kinetics of Porcine Alpha Amylase. Presented at the Research Congress 2013. De La Salle University Manila. 5 pp.
8. Güdr A. (2016). Influence of Total Anthocyanins from Bitter Melon (*Momordica charantia* Linn.) as Antidiabetic and Radical Scavenging Agents. *Iranian journal of pharmaceutical research : IJPR*, 15(1), 301–309.
9. Hasan, R., Uddin, N., Hossain, M., Hasan, M., Yousuf, E., S, L., . . . Choudhuri, M. (2014). In vitro  $\alpha$ -amylase inhibitory activity and in vivo hypoglycemic effect of ethyl acetate extract of *Mallotus repandus* (Willd.) Muell. stem in rat model. *Journal of Coastal Life Medicine*, 2(9), 721-726. doi:10.12980/jclm.2.201414d29
10. He, K., Song, S., Zou, Z., Feng, M., Wang, D., Wang, Y., Li, X., and Ye, X. 2016. The Hypoglycemic and Synergistic Effect of Loganin, Morroniside, and Ursolic Acid Isolated from the Fruits of *Cornus officinalis*. *Phytotherapy Research* 30 (2): 283-291.
11. Heralde III, F.M., Nicodemus, N. Yu, G.F., and Juinio, H. 2017. Evaluation of the Safety and Health Potentials of Bawang and Ampalaya in Hypertension and Type 2 DM through Metabolic Profiling/Metabolomics. Unpublished.
12. Liu, S., Yu, Z., Zhu, H., Zhang, W., and Chen, Y. 2016. In vitro  $\alpha$ -glucosidase inhibitory activity of isolated fractions from water extract of Qingzhuang dark tea. *BMC Complementary and Alternative Medicine* 16, 378.
13. Nair, S. S., Kavrekar, V., & Mishra, A. (2013). In vitro studies on alpha amylase and alpha glucosidase inhibitory activities of selected plant extracts. *European Journal of Experimental Biology*, 3(1), 128-132.
14. Pulbutr, P., Saweeram, N., Ittisan, T., Intrama, H., Jaruchotik, A., & Cushnie, B. (2017). In vitro  $\alpha$ -amylase and  $\alpha$ -glucosidase Inhibitory Activities of



*Coccinia grandis* Aqueous Leaf and Stem Extracts. *Journal of Biological Sciences*, 17(2), 61-68. doi:10.3923/jbs.2017.61.68

15. Riaz, T., Abbasi, M., Shahzadi, A., Ajaib, M., and Khan, k. 2012. Phytochemical screening, free radical scavenging, antioxidant activity and phenolic content of *Dodonea viscosa* Jacq. *J. Serb. Chem. Soc.* 77 (4): 423-435.
16. Wang, L., Clardy, A., Hui, D., Gao, A., and Wu, Y. 2019. Antioxidant and antidiabetic properties of Chinese and Indian bitter melons (*Momordica charantia* L.). *Food Bioscience* 29: 73-80.
17. <https://starnage.com/app/global/campaign/sampling-detail/209>

### Supplemental Table and Figures.

Table S1. Coordinate settings used in the in silico docking simulations.

Binding Site	Allosteric Site 1		Allosteric Site 2		Allosteric Site 3		Active Site	
	C	D	C	D	C	D	C	D
X	37.660	21.059	27.015	22.636	40.478	16.269	36.302	11.109
Y	30.372	30.117	15.855	16.615	47.721	16.749	36.178	16.892
Z	30.000	25.090	16.615	20.007	22.862	19.965	4.925	19.504

Legend: C = center; D = dimension

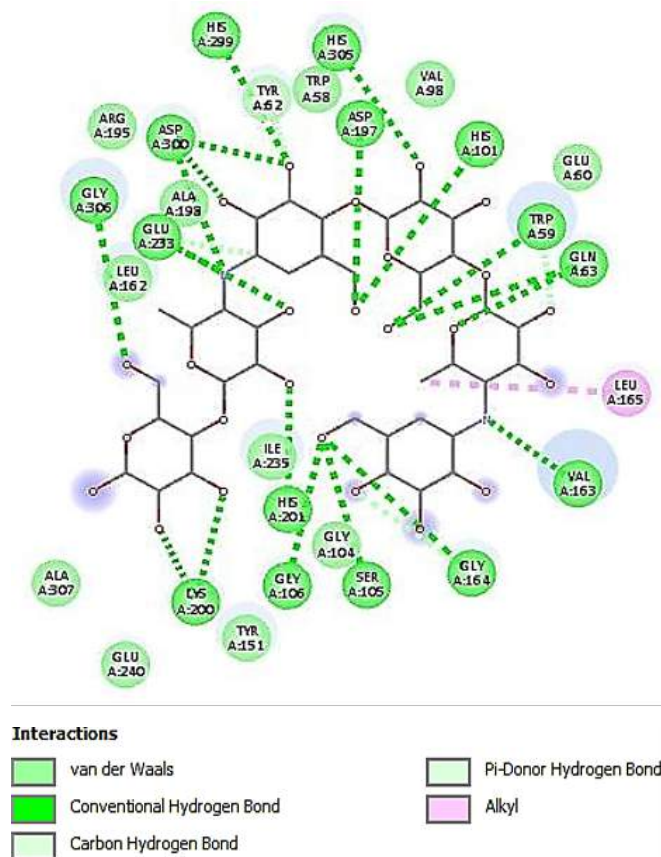


Figure S1. Interaction of acarbose with the residues present in the active site of 10SE.

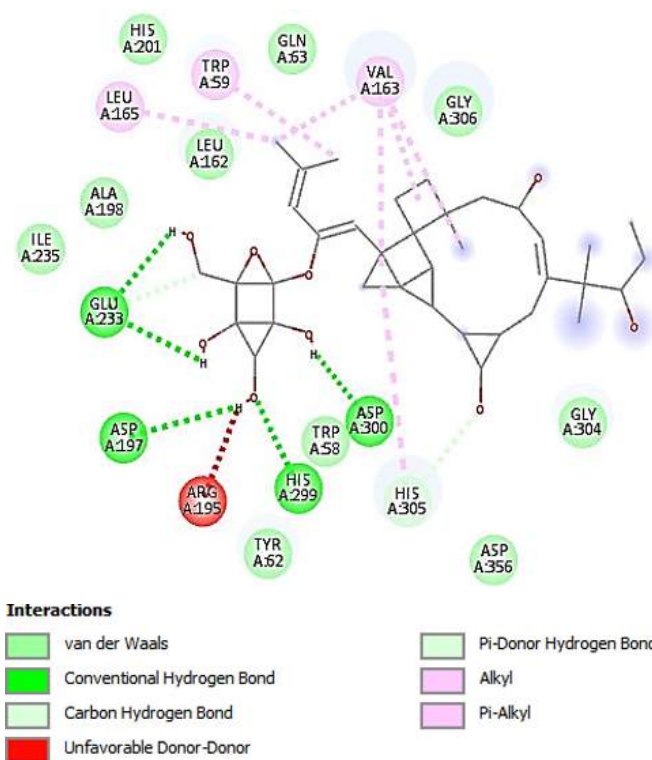


Figure S2. Interaction of 1889 compound with the residues present in the active site of 10SE.

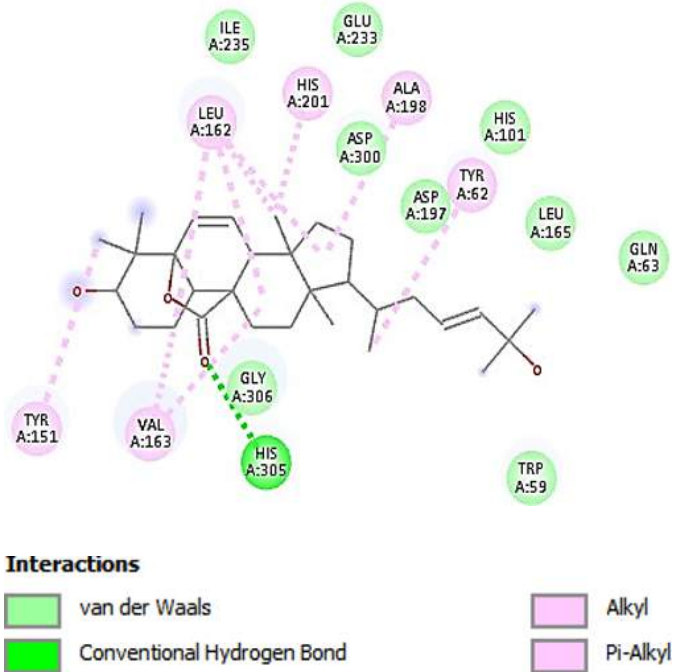
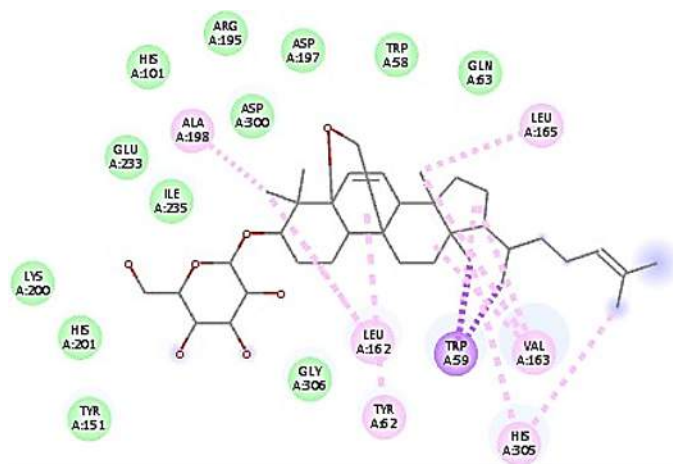


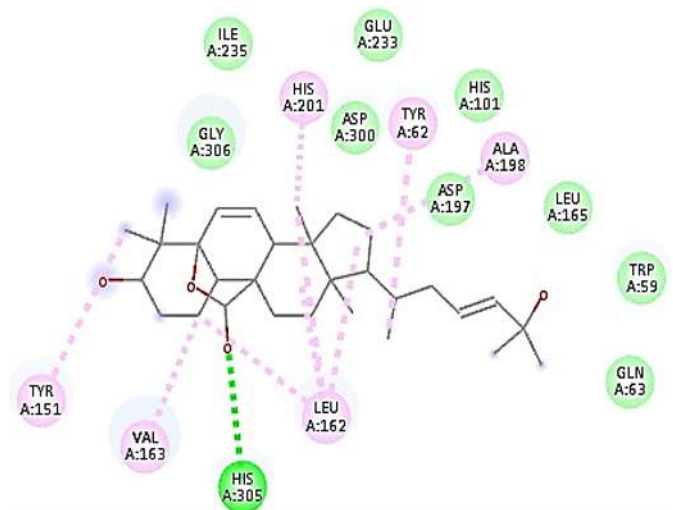
Figure S3. Interaction of Compound 0179 with the residues present in the active site of 10SE.



### Interactions



Figure S4. Interaction of Compound 6169 with the residues present in the active site of 1OSE.



### Interactions

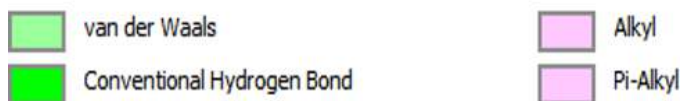
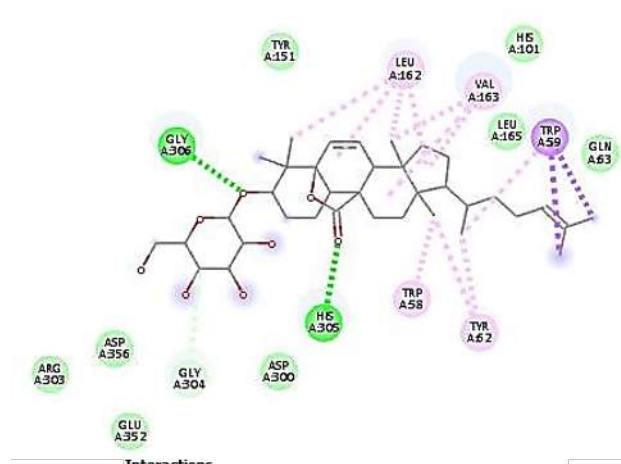


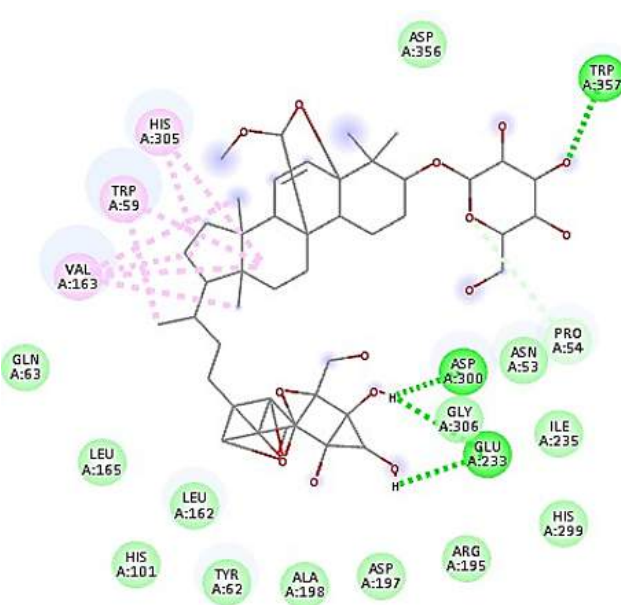
Figure S5. Interaction of Compound 8554 with the residues present in the active site of 1OSE.



### Interactions



Figure S6. Interaction of Compound 6324 with the residues present in the active site of 1OSE.



### Interactions

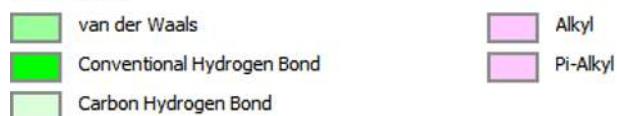


Figure S7. Interaction of Compound 2341 with the residues present in the active site of 1OSE, when 8241, 2339 and 2341 are already bound in the allosteric sites.

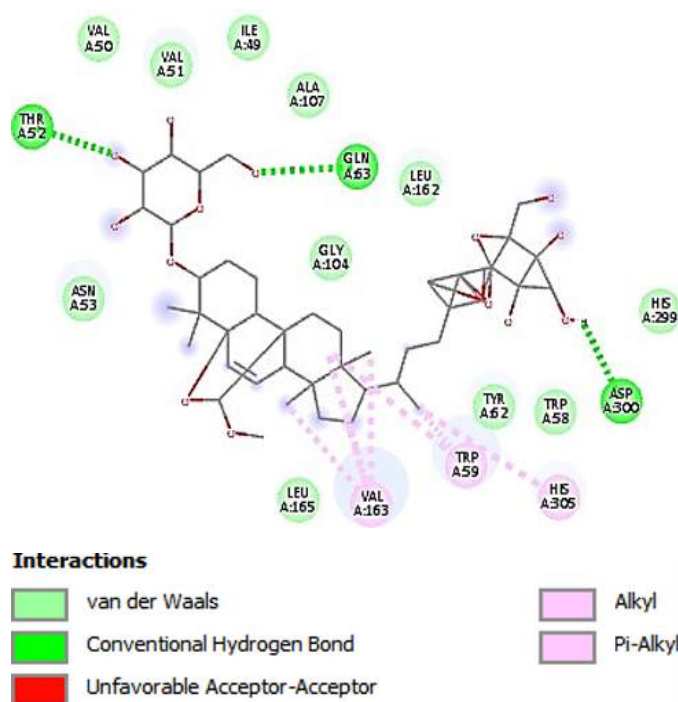


Figure S8. Interaction of Compound 2341 with the residues present in the active site of 1OSE, when 6269, 6079 and 2341 are already bound in the allosteric sites.

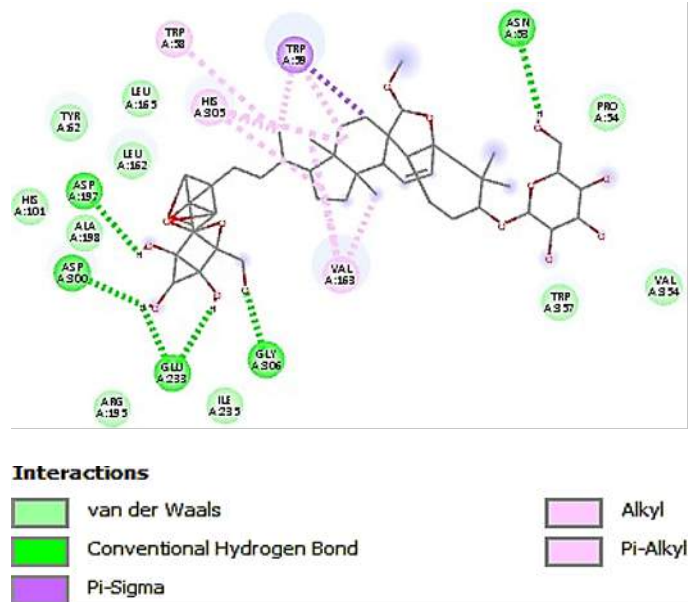


Figure S9. Interaction of Compound 2341 with the residues present in the active site of 1OSE, when 6269, 2341 and 2341 are already bound in the allosteric sites.



January-December 2020 | Volume 1, Issue 1&2

### **Guide to Contributors**

Electronic copy of the manuscript should be addressed to the Editor-in-Chief, PJBMB, care of the Department of Biochemistry and Molecular Biology, U.P. College of Medicine, P.O. Box 593, Manila.

### **Manuscript Preparation**

**Typing.** Manuscript should be encoded, double-spaced, using the font Arial, size 12, with a 2.5 cm margin around. When applicable, each selection should begin on a new page (title page, abstract, introduction, methods, results, acknowledgements, references, tables, figures). Each page should be numbered, starting with the title page.

**General Rules on Style.** All symbols, abbreviations, and acronyms should be defined. All acknowledgements should be gathered into a brief statement at the end of the references and notes. Tables should supplement, not duplicate, the text. They should be numbered according to the order of their citation in the text.

**Title Page.** This page should contain the title of the manuscript, names, addresses, and telephone numbers of the authors, and the laboratory where the work was done. The author who is responsible for correspondence should be indicated. Titles of general articles should have no more than 50 characters.

**References and Notes.** References should be listed and alphabetically numbered accordingly. Conventional abbreviations for journals should be used

### **Categories of Articles**

There are seven categories of papers that are published.

**Biochemical Education Articles.** A biochemical education article (up to 3000 words) is expected to describe a personal approach to teaching a concept, or to review current trends in the teaching of biochemistry. A class laboratory experiment that students have found especially helpful may be submitted. The use of illustration is encouraged, but should be limited to three. Figures and tables together should not exceed four. References should be included when applicable, but are limited to 20.

**Research News.** A research news article (up to 250 words) is expected to describe recent (not older than six months) work in biochemistry and related fields. Abstracts may be submitted as news. The use of drafts and tables is discouraged.

**Research Articles.** A research article (up to 4000 words) is expected to contain new data in its field. The article should include an author note (name, title and address), abstract, introduction, materials and methods, results and discussion, references and notes. The introduction should outline the main point of the paper and should not exceed 150 words. A maximum of 30 references is suggested. Figures and tables together should not exceed six.

**General Interest Articles.** General interest articles (up to 5000 words) are expected to describe developments that do not fit in the research or education categories, but may have biochemical applications. General interest articles should include the author's name, address, and title, a summary (up to 100 words) that outlines the main points of the articles; and brief subheadings to highlight main ideas. Figures, tables, and sharp black and white photographs and cartoons may be submitted.

**How-To Articles.** A how-to article (up to 2000 words) is expected to provide step-by-step instruction on useful activities relating to biochemistry. Illustrations are encouraged but should be limited to four.

**Letters.** Letters to the editor that discuss topics published in the PJBMB will be considered for publication. Such letters may correct errors, reinforce ideas, or provide alternative perspectives. When the letter cites errors, the author of the PJBMB will be given a chance to reply. Letters should not exceed 150 words.

**Book Reviews.** A book review (up to 500 words) is expected to compare the book with others of its kind and suggest for whom it may be valuable. The review should include a publisher's note (title, author, publisher, year of publication, number of pages, price), and overview of its contents, features that set it from the rest of its kind, or that make it worthwhile reading, and conclusion as to whether it is recommended or not.