Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

EXOME DATA ANALYSIS IN THE DISCOVERY OF VARIANTS ASSOCIATED WITH BREAST CANCER METASTASIS AND THEIR IMPLICATIONS ON PROTEIN STRUCTURE

Jaswanth Jenny P[1], Dhamotharan R[2]

1. Department of Plant Biology and Plant Biotechnology, Presidency College, Chennai (TN), India
2. Associate Professor, Department of Botany, Government Arts College of Men (Autonomous), Chennai (TN), India

ABSTRACT

Understanding the correlation between heterogeneity in terms of genomic alterations and metastatic behavior is crucial in improving outcomes for breast cancer patients. Exome sequence that encode proteins generally encompasses around 2% of the genome, and is believed to harbor up to 85% of all disease causing variants. In this study, whole exome data from primary, metastatic and matched normal breast cancer samples submitted by Memorial Sloan Kettering Cancer Center (PRJNA273304) was deep dived using GATK pipeline, specifically Haplotypecaller was employed for variant calling and filtered using annovar. Impact of the identified mutations on the structure of the proteins was also analyzed. Variation in TOGARAM2, C3orf18, S100Z, MYH15, EPB41L4A, LARP1B, NAALADL2, OR2W3 and OR2AK2 were found in all nine primary and metastatic samples. None of these variants were previously reported to be associated with *any* disease conditions. Based on the structural availability of templates, MODELLER was utilized to build the three-dimensional structures of S100Z, MYH15, NAALADL2, OR2W3 and OR2AK2, and validated using the web based PROCHECK server. Lastly, molecular dynamics was carried out to evaluate the stability of the modeled proteins. Except OR2W3, all other mutant proteins exhibited significant RMSD deviation in the simulation studies substantiating the role of mutations. We conclude that the mutations identified can be useful in recognizing patients with breast cancer who are likely to develop remote metastases.

KEYWORDS:

Exome sequencing, GATK, Molecular modeling, Molecular dynamics, Breast cancer metastatic

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

INTRODUCTION

Breast cancer is the most frequent malignancy affecting women, and WHO reports 6,27,000 deaths caused by breast cancer in 2018. Approximately 90% of cancer-related mortality are caused by metastases or secondary tumorsin distant locations from the primary tumor.(1) Multiple genes and biomolecules contribute to the complex mechanism of breast cancer metastasis.(2) Metastasis cascade involves angiogenesis, detachment of primary tumorcells, migration, intravasation, adhesion of tumorcells in a distant organ site, extravasation followed by colonization.(3) Metastatic breast cancers very rarely display clinical symptoms and signs.(4, 5)Genetic mutations are presumably believed to be accumulated progressively in surviving cancer cells and expend through tumorclonal evolution. Studies mention that a small cell subpopulation of cells gets separated with acquired metastatic capacities within the primary tumorand is very effective in colonizing distant organs, suggesting that metastasis-driving mutations occur frequently in the distant metastases than in the primary tumor.(6) Numerous studies have reported mutations and mutational signatures responsible for distant metastases. Long-term survival of cancer patients can be improved by focusing on prevention and treatment of metastatic disease by specifically blocking the colonization of secondary organs.(7) Dissemination of tumorcells in the later stage of cancer that seed metastasis or local relapse suggests that the primary tumorgenome hold the information to predict the probability of occurrence of metastasis.(8)

Exome represents only around 2% of the human genome that hold information regarding 85% of known disease-causing variants thereby making exome sequencing cost-effective and potential approach for finding disease genes.(9)It acts as an efficient tool in detecting the genetic basis of diseases and traits that have proved to be unmanageable to traditional gene-discovery strategies. Additionally, with the advent of exome sequencing rare alleles may explain the heritability of complex diseases, evaluate disease risks and health-related traits.(10) Whole exome sequencing has proven to be a powerful tool for facilitating clinical diagnosis and deciding personalized treatments. Applying exome sequencing in single proband or multiple diseased individuals have demonstrated tremendous success in identifying disease mutations.(11) Exome sequence analysis is recognized as a cheap and quick genetic diagnosis and has paved the way for identification of novel disease-causing mutations and promising new therapies. Projects like Cancer Genome Atlas (TCGA), International Cancer Genome Consortium and 1000 Genomes project have facilitated the discovery of major cancer-causing genomic alterations and candidate drug targets.(12)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

Various studies have deciphered on the role of the mutations leading metastasis of breast cancer. *Celine Lefebvre et.al.,* worked on 216 breast cancer tumor-blood pairs and suggested that eight genes (*ESR1*, *FSIP2*, *FRAS1*, *OSBPL3*, *EDC4*, *PALB2*, *IGFN1*, and *AGRN*) were found to be more frequently mutated in metastatic breast cancer compared to primary tumor.(13)*Aravind Kumar M et. al.,* identified a metastasis related variant, MMP9 rs199676062 in triple negative breast cancer cases.(14)

DNA mutations occurs through different mechanisms and mostly create a prominent impact on biological function while few are relatively 'silent'. Analyzing the influence of mutation on protein structure stability has huge potential to guide pharmaceutical drug design initiatives which aim to prevent the effects of deadly diseases. (15) Proteins play an imperative role in living organisms, and studies have proved their positive correlation with the protein structure. Structural data of proteins provide the basis for rationalizing correlation between constituent molecules and specific functional evident processes. (16) Mutagenesis studies of physical proteins in wet-lab involves required laboratory setup, time and cost, whereas computational studies compliment wet-lab and also provides the feasibility of exploring various approaches with promising accuracy rates.

Emergence of faster and more powerful computers have explored more complex systems using computer modelling or computer simulations. Stability of the protein structure depends upon the environment. Molecular dynamics studies help us to know impact of environment on protein structure. Theoretically, rate of catalysis is demonstrated by binding variations between the ground and transition state. (17) Molecular motions play a vital role in the conformation of the protein structure, thereby creating an impact in the function of the protein.

In this study, whole exome sequence data of nine breast cancer patients affected with metastasis was extensively studied by applying GATK-pipeline. Exome sequence of paired normal, synchronous primary and metastasis of invasive breast cancer was analyzed for novel mutations responsible for metastatic breast cancer. Genetic variants were predicted using GATK pipeline best practices. Additionally, molecular modeling and dynamics studies were performed to infer the likely contribution of driver mutations in breast cancer metastasis that were shared among all metastases in each patient.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

MATERIALS AND METHODS

Data source: Exome sequence of paired normal, synchronous primary and metastasis of invasive breast cancer holding the accession number PRJNA273304 submitted by Memorial Sloan Kettering Cancer Center were obtained from NCBI – Sequence Read Archive (SRA) (SRP055001). PRJNA273304 holds 27 samples, three pairs of matched normal, primary and metastasis breast cancers. (18) These exome data are from diagnostic biopsies of nine patients with primary tumors and synchronous distant metastases presenting stage IV breast cancer before systemic treatment.

Exome sequence analysis pipeline: High throughput sequencers generate hundreds of millions of sequences in a single run. Before processing, some simple quality control checks should be performed on the sequences to draw biological conclusions. FastQC was used to generate a QC report as a part of analysis pipeline that helps on identifying problems originated either in the sequencer or in the starting library material. MultiQC was employed to consolidate the FastQC results to collectively visualize the output across the samples analyzed, thereby enabling to understand the global trends and biases are identified quickly. Fastq format of the exome sequences were converted to Fastqsanger using FastQ Groomer. Trimmomatic is incorporated into the pipeline to trim low-quality bases from reads and remove low-quality reads. Resulting sequence reads were mapped with genome using Bowtie2. Latest version of GATK Haplotype caller (HC) was used to call variants from the whole exome sequence. Pair Hidden Markov Model (HMM)s forward algorithm efficiently accelerates the GATK HC execution. (19) Variants called using GATK-HC was annotated using Annovar. It is a command-line driven software tool can be used as a standalone application on diverse hardware systems where standard Perl modules are installed.

Pfam data warehouse of protein families, which is classified into manually curated Pfam-A and automatically generated Pfam-B families. (20) Dynamic programming algorithm Basic local alignment search tool (BLAST) against the Protein Data Bank (PDB) was used to find the template. Phyre2 server was used as an alternative to gather homologous sequences by searching against the specially curated nr20 protein sequence database with HHblits. Tertiary structure of the protein sequences with 40% identity covering the domain region were predicted using Modeller9.21 package. (21) Template structure with 40% identity results in a model structure with good stereochemistry and RMSD value less than 1 Å. Alignment with template sequence was performed using malign.py program and the homology model of the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

protein was built using model-default.py. Modelled structure was evaluated using TM-align server, which is an algorithm that identifies the best structural alignment between the target and template sequence. Overall quality of the modelled protein structure was evaluated by plotting Ramachandran Plot using Procheck Server.(22)

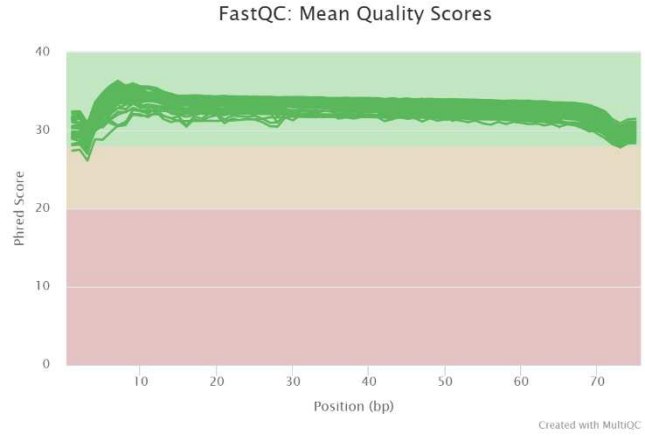
Molecular dynamics studies: Molecular dynamics simulations were performed using GROMACS for 20 nanoseconds (ns). The simulation system was prepared by addition of hydrogen atoms to the .pdb coordinates, placed in dodecahedron box followed by solvation. Prior to simulation, the system was relaxed with energy minimization to make sure there is no steric clashes or inappropriate geometry within the system. Unrestrained dynamics leads to collapse of the system, to avoid this solvent and ions around the macromolecule should be equilibrated. Two different phases are involved in equilibration. The **first phase** in conducted under NVT ensemble (constant Number of Particles, Volume and temperature). grompp and mdrun were invoked as for energy minimization. **Second phase** of equilibration in performed under an NPT ensemble, where the number of particles, pressure and temperature are all constant. After achieving the desired temperature and pressure, the position restraints will be released, and molecular dynamics simulation was performed. The obtained trajectories were evaluated for Root Mean Square Deviation (RMSD) to understand the convergence of the macromolecule happened during the simulation period, Radius of Gyration (Rg) to compute the radius of gyration for the whole macromolecule studied and Root Mean Square Fluctuation (RMSF) to measure the flexibility of an amino acid.

RESULTS

Quality control of Exome data: The metagenome comprised of 54 raw reads with a G + C content of 47%. More than 85% of the sequence was having a Phred score>=Q30, which represents the quality of the sequence as 99.90%. Most of the quality calls on platforms will be reduced as the run progresses, so it is common to visualize few base calls falling into the region of reasonable quality towards the end of the read.

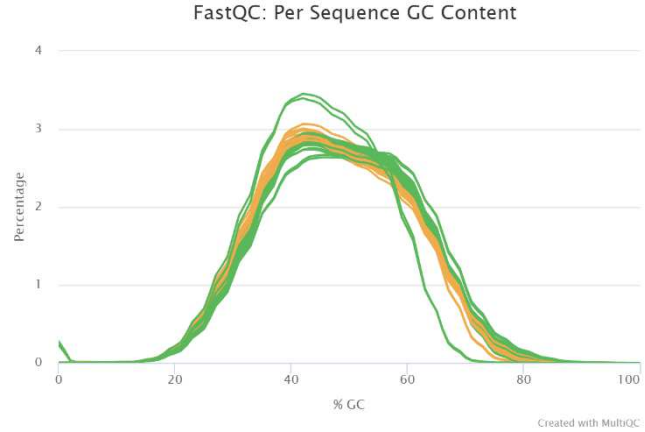
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

Fig. 1

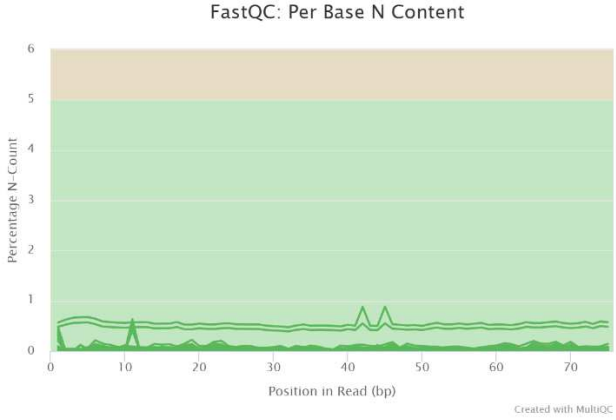


(a)

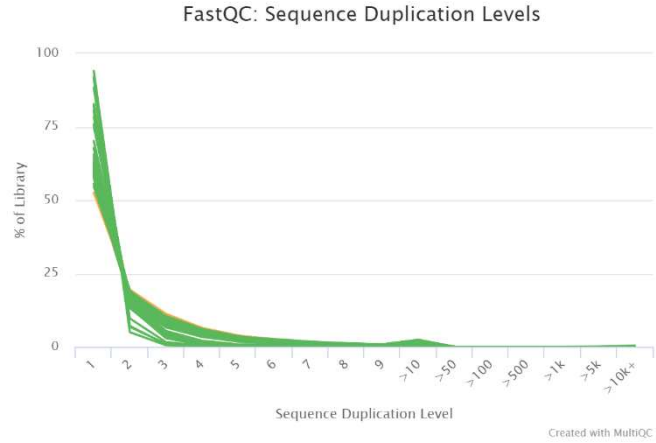
(b)



(c)



(d)



(e)

Figure. 1(a):Y-axis represents the quality scores. Green - Very good quality calls; Orange – Calls of reasonable quality; Red – Calls of poor quality. Majority of the sequence was having

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

a phred score ~99.90%**(b):**Per sequence quality scores shows an error rate below 0.2% **(c):**Thirty-three (33) samples passed, whereas warning was raised for 21 samples. **(d):**Below 1% percentage of base calls at each position were found to be substituted by N. **(e):**Sequence duplication levels were less than 20%.

The per sequence quality score report let us know the sequence reads were below the error rate of 0.2%. The central peak of Fig. 1 (c) corresponds to the overall GC content of the studied exomes. Thirty-three samples passed this criterion, whereas warning was raised for 21 samples (i.e., sum of the deviations from the normal distribution is more than 15% of the reads). Per base sequence content was found to be less than 10% for all samples, except SRR1802852 (i.e., the difference between A and T, or G and C is greater than 10% in any position). Per Base N Content graph plotted suggested less than 1% of base calls at each position were substituted by N. All samples had sequences of a single length (75bp). Sequence duplication levels were found to be less than 20%, except for SRR1802836_1. All 54 samples had less than 1% of reads made up of overrepresented sequences, whereas no samples shown adapter contamination > 0.1%.

Pre-processing of exome data: Exome sequence data retrieved from NCBI-Sequence Read Archive (SRA) in fastq format with quality values was changed to Sanger-conforming fastq format using FASTQ groomer, so that it can be used in downstream application such as mapping. Trimming tasks for illuminapaired end and single ended data, such as adapter sequences and low quality regions were performed using trimmomatic. Short trimmed sequence of each dataset was aligned to human genome hg19 (GRCh37 Genome Reference Consortium Human Reference 37) using Bowtie2.

Variant Calling and Annotation: Single nucleotide variants (SNV) were called with the Genome Analysis Toolkit HaplotypeCaller. Output from variant calling was directly used for SNV detection by Annovar. Gene based annotation of all the variants generated by HaplotypeCaller was performed. Nine variants indicated as damaging by Sorting intolerant from tolerant (SIFT) and Polphen2 that are common in metastatic and primary samples were selected for further analysis, while all these variants were predicted as harmless (Polymorphism automatic) by MutationTaster. (Table 2) LRT algorithm indicates variants in C3orf18, S100Z, NAALADL2, OR2AK2, LARP1B and OR2W3 as deleterious variants.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

Table. 2: lists the chromosomal positions in which variants are identified in all nine respective samples.

5736832T>C	
6199984T>A	
6328947T>C	
10955175C>T	
29003646A>G	
31349841C>T	
50553661G>A	
59792934G>T	
61125763C>T	
76875427A>C	
108470146G>A	
112168782C>T	
128122049C>G	
175627355C>G	
247606450A>G	
247856174A>C	
247932938G>A	
6328947T>C	

only in primary
Mutation in primary and metastasis
Mutation in primary, metastasis and normal
Only in normal

Mutations shared in both the metastatic sample and the primary tumorgiving rise to the metastasis are indicated in green, mutations only seen in the metastatic sample are indicated in blue, mutations identified in all three samples are indicated in red and mutations only observed in the normal samples are indicated in purple.

A total of 17 variants were found to be common in all primary samples, 9 variants in metastatic samples and 2 variants in normal samples. Mutations identified to be common in primary and metastatic samples were never previously described in literature. Nine Variants listed in Table. 3 exhibited concordance among primary and metastatic samples, specifically OR2AK2 c.G607A (p.V203M) was observed in normal, primary and metastatic samples. All the short-listed variants are represented in dbSNP and Exome Aggregation Consortium (ExAC) database. None of the variants were reported in ClinVar, which shows no association of these variants with disease or clinical phenotype are reported.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

Table. 3: Candidate variants selected for further analysis

Gene	NM#	Exon #	Nucleotide change	Aminoacid change	Chr Position	Zygosity
TOGARAM2	NM_199280	exon6	c.A794G	p.Q265R	29003646	hom
C3orf18	NM_016210	exon6	c.C485T	p.A162V	50553661	hom
S100Z	NM_130772	exon3	c.A68C	p.E23A	76875427	het
MYH15	NM_014981	exon15	c.C1510T	p.H504Y	108470146	het
EPB41L4A	NM_022140	exon22	c.G1889A	p.R630H	112168782	hom
LARP1B	NM_018078	exon11	c.C1385G	p.P462R	128122049	hom
NAALADL2	NM_207015	exon11	c.C1865G	p.P622R	175627355	het
OR2W3	NM_001001957	exon1	c.A588C	p.E196D	247856174	hom
OR2AK2	NM_001004491	exon1	c.G607A	p.V203M	247932938	hom

Similarly, none of the genes with the variants were listed in Drug Gene Interaction data (http://dgidb.genome.wustl.edu/).

Shortlisted mutations from exome analysis were validated for its stability and deleterious effects using i-Mutant 3.0 and PROVEAN. S100Z E23A, LARP1B P462R, NAALADL2 P622R and OR2W3 E196D mutations were found to cause deleterious effects.

Gene	AA Change	i-Mutant 3.0		PROVEAN	
		Stability Prediction	Reliability Index (RI)	PROVEAN Score	Prediction (Cutoff = -2.5)
TOGARAM2	Q265R	Neutral	7	-1.337	Neutral
C3orf18	A162V	Neutral	1	-0.575	Neutral
S100Z	E23A	Disease related	2	-5.302	Deleterious
MYH15	H504Y	Neutral	0	-4.591	Deleterious
EPB41L4A	R630H	Neutral	2	-0.093	Neutral
LARP1B	P462R	Disease related	5	-4.129	Deleterious
NAALADL2	P622R	Disease related	6	-5.623	Deleterious
OR2W3	E196D	Disease related	4	-2.834	Deleterious
OR2AK2	V203M	Disease related	1	-1.052	Neutral

Domain analysis and molecular modelling: Domain analysis of S100Z demonstrated significant matches with S-100/ICaBP type calcium binding domain, which belongs to pfam

(Clan) EF-hand like superfamily comprising 31 members. Glu at position 23 was manually changed to Ala and the structures was modelled using modeller9.21. Ramachandran plot suggested 95.7% residues in the allowed region and the RMSD score between the wild type and mutated structure was 0.36 Å. Results shows the mutation is identified in the hinge region of EF-hand domain of S100Z.

Pfam analysis of MYH15 was found to be matching with Myosin N-terminal SH3-like domain (Myosin N) and Myosin head (motor domain) and coiled coil, Myosin tail 1 domains. Homology search against PDB database PDB ID: 5H53 with 64% identity. Interestingly, Myosin N and Myosin head domains were found to be within the BLAST alignment suggesting the appropriate selection of 5H53 as a template for homology modelling. Domain (106 – 778) was modelled using modeller9.21 and the modelled structure was validated. Procheck generated Ramachandran plot shows the presence of 89.9% residues in the most favoured regions, and very minimal residues (1%) were in disallowed regions. TM-align computed RMSD value of 0.53 Å suggests that modelled MYH15 is of good quality. RMSD values computed for wild type and mutant model was 0.25 Å, suggesting very low impact of the H504Y mutation in the structure.

NAALADL2demonstrated the presence of Peptidase Family M28 in pfam analysis and showed 27% identity with PBD ID: 1Z8L. In-spite of not having the 40% identity to be homology modelling template, NAALADL2 was modelled using 1Z8L as template since the domain region was present within the BLAST alignment. Secondary structure prediction using PDBsum showed an addition of beta sheet in the mutant protein. Procheck generated Ramachandran plot showed 82.3% residues of the wildtype and 83.2% of the residues of mutant structure in the most favored region, suggesting an average quality model. RMSD between the template and NAALADL2 modelled structure was 2.02 Å, whereas the RMS deviation between the mutant and the wild-type structure was 0.53 Å.

Primary sequence of OR2W3 and OR2AK2was found to be aligning with 7-tm_4 family of transmembrane olfactory receptors Pfam domain. Phyre2 server suggested PDB ID: 4ZWJ (human rhodopsin) as a potential template with 21% identity covering 98% of the target OR2W3 sequence and 14% identity covering 92% sequence coverage of the OR2AK2 sequence. Protein structure of OR2W3 and OR2AK2 were modelled using modeler 9.2.1 with PDB ID: 4ZWJ as template. Structural analysis did not show any significant difference between the mutant and the wild-type modelled structures.TM-align calculated RMSD between OR2AK2 wild-type and mutant (V203M) was 0.22 Å.

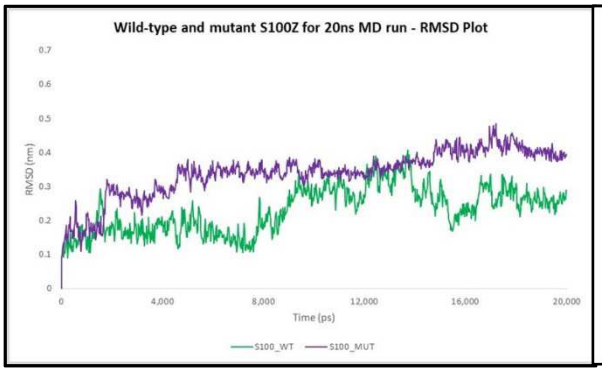
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

TOGARAM2, C3orf18, LARP1B and EPB41L4A were ignored for the rest of the study for multiple reasons.

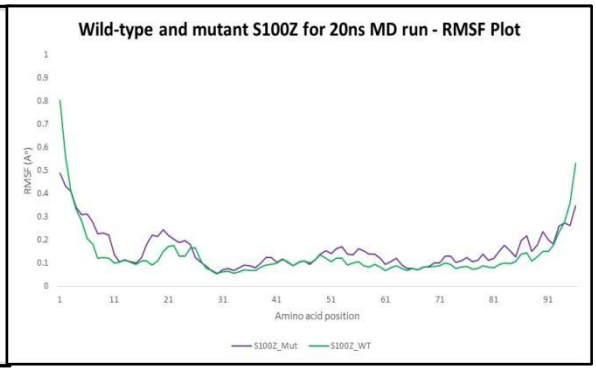
Molecular dynamics studies: Five of the wild-type and mutant modelled proteins, S100Z, MYH15, NAALADL2, OR2AK2 and OR2W3 were analyzed with simulation studies using GROMACS v5.0.5 and the GROMOS96 43a1 forcefield. Simulation was performed for 20ns and the total system was filled with tip3p water model. RMSD of the wildtype structure was considered as the benchmark for measuring the system.

S100Z mutant (p.E23A) protein exhibited distinct RMSD deviation from wild type approximately between 3ns to 10ns with the RMSD more than 0.3 nm and after ~14 ns to the end of the simulation. Figure. 1 (a) RMSF value at the mutant position Glu 23 was found to be 0.17 and 0.2 for the mutant and the wildtype, respectively. Figure. 1 (b) RMSF of residues in S100Z mutant protein fluctuates between 0.0531 to 0.4886 with mean of 0.1565, whereas in S100Z wild-type 0.0565 to 0.8042 with mean of 0.1334. The mutant protein exhibited compact conformation with lower values of Rg which were between ~1.4 and ~1.35 nm. Hydrogen bond interaction between Gly 21 and Arg 24 in the wild-type structure, was not observed in the mutant S100Z structure.

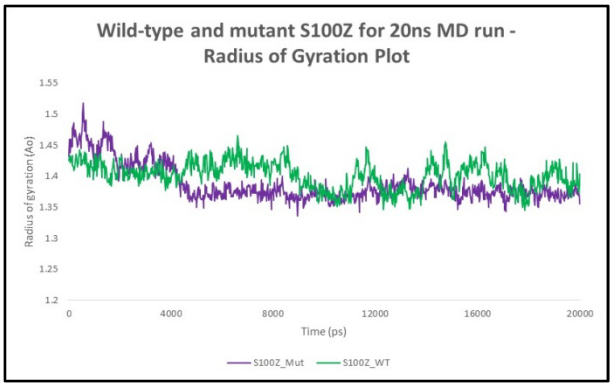
Figure. 1 Comparison of RMSD, RMSF and Rg values between wild-type and mutant S100Z



(a)



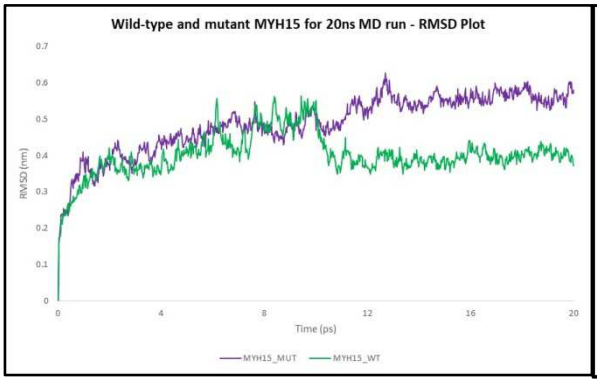
(b)



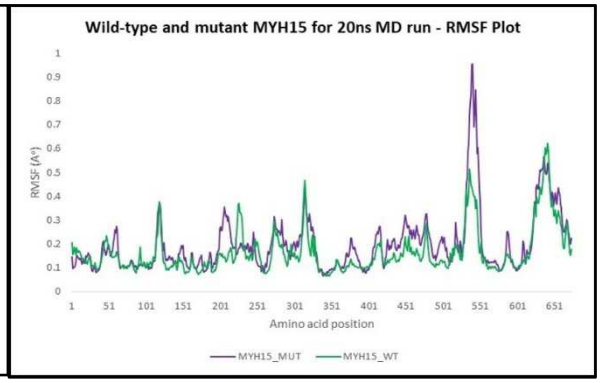
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

(c)

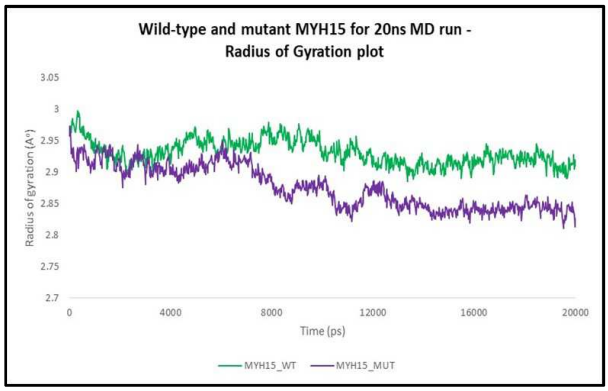
Figure. 3Comparison of RMSD, RMSF and Rg values between wild-type and mutant MYH15



(a)



(b)



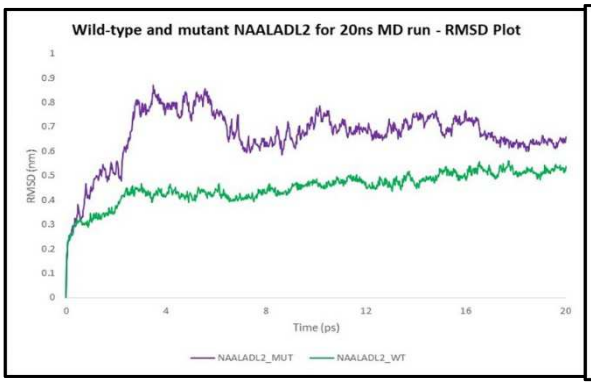
(c)

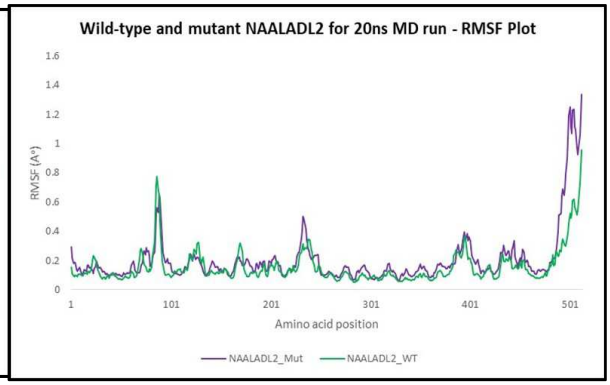
MYH15 mutant (p.H504Y) demonstrated significant RMSD variation between the wild-type and the mutant was noted after ~10 ns (0.55 and 0.47, respectively) that extended till the end of the simulation. Figure. 3(a). Significant RMSF deviation was noted from the amino acids at position 535 to 540 ranging from 0.42 and 0.62 (WT and MUT), and 0.42 and 0.96, respectively. Figure.3(b) Rg plot depicted higher values of wild-type structure throughout the simulation. Remarkable variation was noted approximately after ~7 ns. Figure. 3(c) Analysis of hydrogen bond interaction showed that the H399 interacting with F395 in wild type whereas the mutant Y399 shifted interaction to L403.

Simulation studies of NAALADL2 (p.P622R) protein demonstrated significant variation in RMSD value approximately after 1 ns that lasted till the end of the simulation. Figure. 4 (a) RMSF values of the C-terminal residues showed significant variation, whereas the RMSF value of the mutated position 457 was found to be 0.14 and 0.21 for wild-type and mutant, respectively. Figure. 4 (b) Rg plot showed higher values of mutant structure in the first half of simulation (till ~12ns), and the Rg values were almost similar in rest of the simulation. Figure. 4 (c) Mutated amino acid Arg at 457 was found to interact with Ile 452 and His 462.

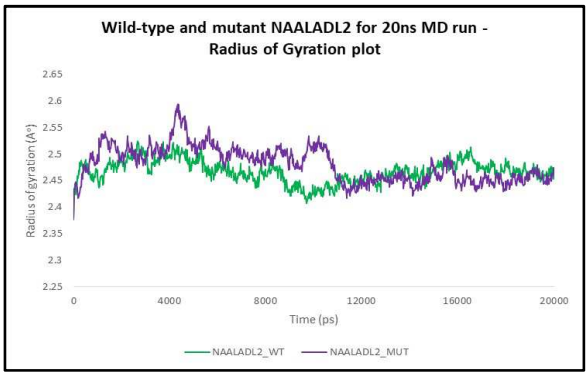
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.

Figure. 4 Comparison of RMSD, RMSF and Rg values between wild-type and mutant NAALADL2





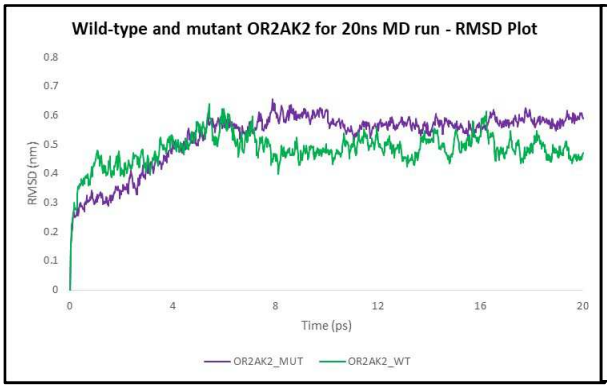
(a) (b)

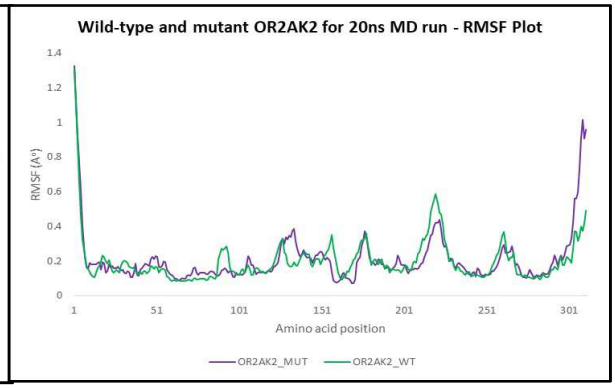


(C)

Molecular dynamics of OR2AK2 wild-type and mutant (V203M) showed a distinct variation in RMSD values after ~1 ns, fluctuation remained till the end of the simulation. Figure. 5(a) RMSF values of the OR2AK2 mutant residues were found to be between 0.0695 Åand 1.3258 Å with average of 0.2093 Å, whereas RMSF values of wild-type OR2AK2 residues were in-between 0.0834 Å and 1.3034 Å with an average of 0.1991 Å. Figure. 5(b) Mutated residue at 203 (185) position deviated from 0.1825 to 0.2125 Å. The plot of the variation of radius of gyration (Rg) showed significant increase in the OR2AK2 mutant structure compared to the wild-type structure. Figure. 5(c)

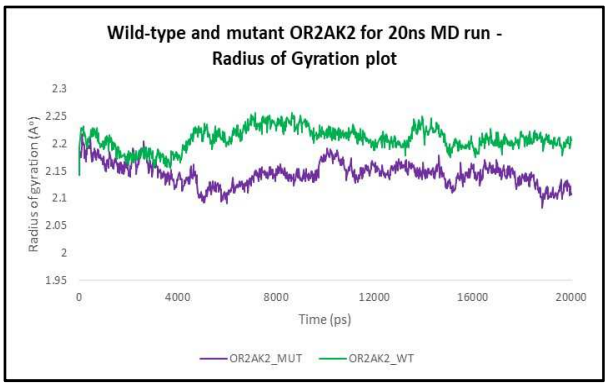
Figure. 5 Comparison of RMSD, RMSF and Rg values between wild-type and mutant OR2AK2





(a) (b)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1663-1682
Received 16 February 2021; Accepted 08 March 2021.



(c)

Simulation studies of OR2W3 did not show significant variations.

DISCUSSION

In this study, a breast cancer exome-wide dataset PRJNA273304 was analyzed for genetic variants responsible for metastasis. Novel potentially targetable alterations were found in the metastases relative to their matched primary tumorsand normal samples. Findings provide evidence to support the contention that studies seeking to define the genetic basis in the metastatic setting ought to perform the genetic analysis rather than the primary tumor.

S100Z is a dimeric, predominantly alpha-helical protein that have demonstrated its ability to bind to calcium ions. (23, 24) S100Z was found to be involved in different stages of breast cancer development.(25)Upregulation of S100Z was observed in metastatic breast cancer tissue compared to normal tissue(25, 26) Glu amino acid at position 23 is one of the Ca2+ coordinate residue in S100Z. Study by PatriziaCancemi et.al., in Affimetrix data (ID 1554876-a-at-at) showed significant association of S100Z expression with relapse free survival and distance metastasis free survival.(27) Yeast-two hybrid study demonstrates the interaction of S100Z with S100P. (23) These Ca2+ sensor proteins regulates its biological activity through transmitting Ca2+ dependent signal to a target protein.(28)*In-vivo* and clinical studies demonstrate a significant role of S100P in breast cancer metastasis in correlation with metastasis-inducing protein, S100A4. (29) Studies suggest binding of calcium with S100Z leads to conformation changes and opens accessibility to hydrophobic surfaces of the protein. Other members of S100 superfamily,S100A4 and S100A14 was found to be significantly associated with induction and tumormetastasis. (30, 31)

Myosin heavy chain 15 (MYH15) protein is detected widely in skeletal muscle, more specifically in fibres of the orbital layer of extraocular muscles and in the extracapsular region of bag fibres.(32, 33) A study in chicken breeds demonstrates involvement of MYH15 in the regulation of muscle growth.(34)EPB41L4A was found to be significantly expressed in

tissues undergoing morphogenetic movements, strongly suggesting a role of EPB41L4A in embryogenesis.(35) Pang, BR et.al., identified that EPB41L4A-AS2 inhibited breast cancer cell proliferation, migration and invasion and induced cell apoptosis in-vitro. (36, 37) Along with its tumorsuppressor role, EPB41L4A was also found to be involved in metabolic reprogramming of cancer by acting as a repressor of the Warburg effect.(38)

N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 (NAALADL2) is a member of the glutamate carboxypeptidase II family that was first identified in a patient with mild mental retardation.(39, 40) Pathogenic variants in NAALADL2 loci was found to be associated with breast cancer risk and Kawasaki disease, prostate cancer and autism spectrum disorder.(41-45)OR2W3 expression is observed in human retinae and olfactory epithelium.(46, 47) Mutations in OR2W3 likely has an association with male reproductive disorders. (48-50) OR2AK2 belongs to the olfactory receptor gene family, which constitute the largest gene family of the mammalian genomes. (51) Nominal association of OR2AK2 mutations with obesity and diabetic and non-diabetic nephropathy has been observed previously. (52, 53) No direct association of OR2AK2 with oncolytic diseases was published till date.

Calculation of RMSD values are the standard tools to find the conformational changes of three-dimensional protein structures. Its significance in comparing protein structure has been evaluated at many instances. (54, 55) Previously, various studies have proved the change in dynamical properties of proteins caused by environmental changes have adapted new functions.(56) The backbone RMSD of the mutant proteins were found to be higher in S100Z, MYH15, NAALADL2 and OR2AK2, and within 1 Å throughout the simulation for all the proteins. Mutant S100Z, MYH15, NAALADL2 and OR2AK2 demonstrated significant RMSD changes throughout the simulation compared to the wild-type structures. No significant changes were observed in OR2W3. Increased in the RMSD values of the mutated S100Z, Myh15, NAALADL2 and OR2AK2 indicates that a significant modification in the overall topology of protein structures. Various studies have proved that perturbations in conformational changes of a protein caused by mutation have paved the way to cancer. (57-59) Allosteric protein modifications can play an imperative role in the cause of cancer and other diseases. (60)

Hydrogen bonds are critical in maintaining the stability and the structural organization in protein structure. (61) Hydrogen bond interaction was not observed in the S100Z mutated residue, but hydrogen bond that existed between Gly 21 and Arg 24 in S100Z wild-type was not observed in the mutated structure. Hydrogen bond interaction of MYH15 His at position 399 with Phe at 395 was found to be shifted in mutant Tyr 399 to Leu 403. Similarly, NAALADL2 mutation at 457 from Pro to Arg changed its interaction to Ile 452 and His 462 from Phe 459 and Asn 460.

The root mean square fluctuation (RMSF) value marks the flexibility of a protein structure which is defined as the distance between atoms and its mean position or the mobility of the atoms pertaining to the average structure throughout the simulation. (62, 63) Over-all, more number of residues of mutant S100Z, MYH15, NAALADL2, OR2AK2 and OR2W3 structures exhibited higher RMSF value compared to the wild-type structure. Increased

RMSD values specifies more flexibility suggesting a role of the mutations in the conformational changes of protein thereby affecting the function of the protein. (64)

Radius of gyration plotted against time would be helpful to understand the degree of compactness and folding of a protein. Degree of compactness is described as the ratio of accessible protein surface area and the ideal sphere of the same volume. (65) All the mutant protein structures analyzed in this simulation study demonstrated flexibility and significant deviation compared to the wild-type protein. Yoon et al., have proved that mutations impact the folding effect on protein structures, which is also evident in this current study. (66) Airy Sanjeev et.al., suggest that increase in Rg values leads to better exposure to solvent and may end in higher aggregation propensity. (67)

The outcomes implicate the link between observed variants and biological phenomena in breast cancer and aids in the discovery of molecular targets for personalized treatment. Molecular modeling and dynamics studies demonstrated that the mutations identified from exome sequencing data may disrupt the protein function and play a role in breast cancer metastasis. These findings support S100Z c.A68C, MYH15 c.C1510T, NAALADL2 c.C1865G, OR2W3 c.A588C and OR2AK2 c.G607A mutations as key drivers, and targeting these genes could be a potential therapeutic direction for metastatic breast cancer.

ACKNOWLEDGMENTS: The first author is grateful to **Memorial Sloan Kettering Cancer Center** for presenting the exome data of subjects of interest in this study in the public resource.

CONFLICT OF INTEREST: None

REFERENCE

1. Spano D, Heck C, De Antonellis P, Christofori G, Zollo M. Molecular networks that regulate cancer metastasis. Seminars in Cancer Biology. 2012;22(3):234-49.
2. Kozlowski J, Kozlowska A, Kocki J. Breast cancer metastasis - insight into selected molecular mechanisms of the phenomenon. Postepy Hig Med Dosw (Online). 2015;69:447-51.
3. Brooks SA, Lomax-Browne HJ, Carter TM, Kinch CE, Hall DMS. Molecular interactions in cancer cell metastasis. Acta Histochemica. 2010;112(1):3-25.
4. Kamby C, Vejborg I, Kristensen B, Olsen LO, Mouridsen HT. METASTATIC PATTERN IN RECURRENT BREAST-CANCER - SPECIAL REFERENCE TO INTRATHORACIC RECURRENCES. Cancer. 1988;62(10):2226-33.
5. Varghese G, Singh SP, Sreela LS. A rare case of breast carcinoma metastasis to mandible and vertebrae. Natl J Maxillofac Surg. 52014. p. 184-7.
6. Chaffer CL, Weinberg RA. A Perspective on Cancer Cell Metastasis. Science. 2011;331(6024):1559-64.
7. Ding L, Ellis MJ, Li SQ, Larson DE, Chen K, Wallis J, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010;464(7291):999-1005.
8. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I, et al. Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell. 2017;32(2):169-+.
9. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011;48(9):580-9.
10. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
11. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics. 2011;12(11):745-55.

12. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1a):A68-77.
13. Lefebvre C, Bachelot T, Filleron T, Pedrero M, Campone M, Soria JC, et al. Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. Plos Medicine. 2016;13(12).
14. Aravind Kumar M, Naushad SM, Narasimgu N, Nagaraju Naik S, Kadali S, Shanker U, et al. Whole exome sequencing of breast cancer (TNBC) cases from India: association of MSH6 and BRIP1 variants with TNBC risk and oxidative DNA damage. Mol Biol Rep. 2018;45(5):1413-9.
15. Dehghanpoor R, Ricks E, Hursh K, Gunderson S, Farhoodi R, Haspel N, et al. Predicting the Effect of Single and Multiple Mutations on Protein Structural Stability. Molecules. 2018;23(2).
16. Lushington GH. Comparative modeling of proteins. Methods Mol Biol. 2015;1215:309-30.
17. Radkiewicz JL, Brooks CL. Protein dynamics in enzymatic catalysis: Exploration of dihydrofolate reductase. Journal of the American Chemical Society. 2000;122(2):225-31.
18. Ng CKY, Bidard FC, Piscuoglio S, Geyer FC, Lim RS, de Bruijn I, et al. Genetic Heterogeneity in Therapy-Naive Synchronous Primary Breast Cancers and Their Metastases. Clinical Cancer Research. 2017;23(15):4402-15.
19. Ren SS, Bertels K, Al-Ars Z. Efficient Acceleration of the Pair-HMMs Forward Algorithm for GATK HaplotypeCaller on Graphics Processing Units. Evolutionary Bioinformatics. 2018;14:12.
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Research. 2012;40(D1):D290-D301.
21. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016;54:5.6.1-5.6.37.
22. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK - A PROGRAM TO CHECK THE STEREOCHEMICAL QUALITY OF PROTEIN STRUCTURES. Journal of Applied Crystallography. 1993;26:283-91.
23. Gribenko AV, Hopper JE, Makhatadze GI. Molecular characterization and tissue distribution of a novel member of the S100 family of EF-hand proteins. Biochemistry. 2001;40(51):15538-48.
24. Streicher WW, Lopez MM, Makhatadze GI. Modulation of quaternary structure of S100 proteins by calcium ions. Biophys Chem. 2010;151(3):181-6.
25. Carlsson H, Petersson S, Enerback C. Cluster analysis of S100 gene expression and genes correlating to psoriasin (S100A7) expression at different stages of breast cancer development. International Journal of Oncology. 2005;27(6):1473-81.
26. Sapkota D, Bruland O, Boe OE, Bakeer H, Elgindi OAA, Vasstrand EN, et al. Expression profile of the S100 gene family members in oral squamous cell carcinomas. Journal of Oral Pathology & Medicine. 2008;37(10):607-15.
27. Cancemi P, Buttacavoli M, Di Cara G, Albanese NN, Bivona S, Pucci-Minafra I, et al. A multiomics analysis of S100 protein family in breast cancer. Oncotarget. 92018. p. 29064-81.
28. Marenholz I, Heizmann CW, Fritz G. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). Biochemical and Biophysical Research Communications. 2004;322(4):1111-22.
29. Wang GZ, Platt-Higgins A, Carroll J, Rudland SD, Winstanley J, Barraclough R, et al. Induction of metastasis by S100P in a rat mammary model and its association with poor survival of breast cancer patients. Cancer Research. 2006;66(2):1199-207.

30. Ambartsumian N, Grigorian M. S100A4, a link between metastasis and inflammation. Molecular Biology. 2016;50(4):510-20.
31. Garrett SC, Varney KM, Weber DJ, Bresnick AR. S100A4, a mediator of metastasis. Journal of Biological Chemistry. 2006;281(2):677-80.
32. Rossi AC, Mammucari C, Argentini C, Reggiani C, Schiaffino S. Two novel/ancient myosins in mammalian skeletal muscles: MYH14/7b and MYH15 are expressed in extraocular muscles and muscle spindles. Journal of Physiology-London. 2010;588(2):353-64.
33. Mascarello F, Toniolo L, Cancellara P, Reggiani C, Maccatrozzo L. Expression and identification of 10 sarcomeric MyHC isoforms in human skeletal muscles of different embryological origin. Diversity and similarity in mammalian species. Annals of Anatomy-Anatomischer Anzeiger. 2016;207:9-20.
34. Zhang ZR, Du HR, Yang CW, Li QY, Qiu MH, Song XY, et al. Comparative transcriptome analysis reveals regulators mediating breast muscle growth and development in three chicken breeds. Animal Biotechnology. 2019;30(3):233-41.
35. Guo YC, Christine KS, Conlon F, Gessert S, Kuhl M. Expression analysis of epb41l4a during Xenopus laevis embryogenesis. Development Genes and Evolution. 2011;221(2):113-9.
36. Pang BR, Wang Q, Ning SP, Wu JQ, Zhang XD, Chen YB, et al. Landscape of tumor suppressor long noncoding RNAs in breast cancer. Journal of Experimental & Clinical Cancer Research. 2019;38:18.
37. Xu SP, Wang PY, You ZL, Meng HX, Mu GN, Bai XN, et al. The long non-coding RNA EPB41L4A-AS2 inhibits tumor proliferation and is associated with favorable prognoses in breast cancer and other solid tumors. Oncotarget. 2016;7(15):20704-17.
38. Liao MJ, Liao WJ, Xu NH, Li B, Liu FH, Zhang SK, et al. LncRNA EPB41L4A-AS1 regulates glycolysis and glutaminolysis by mediating nucleolar translocation of HDAC2. Ebiomedicine. 2019;41:200-13.
39. Whitaker HC, Shiong LL, Kay JD, Gronberg H, Warren AY, Seipel A, et al. N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 is overexpressed in cancer and promotes a pro-migratory and pro-metastatic phenotype. Oncogene. 2014;33(45):5274-87.
40. Borg K, Stankiewicz P, Bocian E, Kruczek A, Obersztyn E, Lupski JR, et al. Molecular analysis of a constitutional complex genome rearrangement with 11 breakpoints involving chromosomes 3, 11, 12, and 21 and a similar to 0.5-Mb submicroscopic deletion in a patient with mild mental retardation. Human Genetics. 2005;118(2):267-75.
41. Walker LC, Marquart L, Pearson JF, Wiggins GAR, O'Mara TA, Parsons MT, et al. Evaluation of copy-number variants as modifiers of breast and ovarian cancer risk for BRCA1 pathogenic variant carriers. European Journal of Human Genetics. 2017;25(4):432-8.
42. Burgner D, Davila S, Breunis WB, Ng SB, Li Y, Bonnard C, et al. A Genome-Wide Association Study Identifies Novel and Functionally Related Susceptibility Loci for Kawasaki Disease. Plos Genetics. 2009;5(1):15.
43. Tonkin ET, Smith M, Eichhorn P, Jones S, Imamwerdi B, Lindsay S, et al. A giant novel gene undergoing extensive alternative splicing is severed by a Cornelia de Lange-associated translocation breakpoint at 3q26.3. Human Genetics. 2004;115(2):139-48.
44. Berndt SI, Wang ZM, Yeager M, Alavanja MC, Albanes D, Amundadottir L, et al. Two susceptibility loci identified for prostate cancer aggressiveness. Nature Communications. 2015;6:7.
45. Kuo PH, Chuang LC, Su MH, Chen CH, Wu JY, Yen CJ, et al. Genome-Wide Association Study for Autism Spectrum Disorder in Taiwanese Han Population. Plos One. 2015;10(9):15.

46. Jovancevic N, Wunderlich KA, Haering C, Flegel C, Massberg D, Weinrich M, et al. Deep Sequencing of the Human Retinae Reveals the Expression of Odorant Receptors. Frontiers in Cellular Neuroscience. 2017;11.
47. Flegel C, Manteniotis S, Osthold S, Hatt H, Gisselmann G. Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing. Plos One. 2013;8(2).
48. Chihara M, Yoshihara K, Ishiguro T, Yokota Y, Adachi S, Okada H, et al. Susceptibility to male infertility: replication study in Japanese men looking for an association with four GWAS-derived loci identified in European men. Journal of Assisted Reproduction and Genetics. 2015;32(6):903-8.
49. Siasi E, Aleyasin A. Four Single Nucleotide Polymorphisms in INSR, SLC6A14, TAS2R38, and OR2W3 Genes in Association with Idiopathic Infertility in Persian Men. Journal of Reproductive Medicine. 2016;61(3-4):145-52.
50. Zhang Y, He XJ, Song B, Ye L, Xie XS, Ruan J, et al. Association of single nucleotide polymorphisms in the USF1, GTF2A1L and OR2W3 genes with non-obstructive azoospermia in the Chinese population. Journal of Assisted Reproduction and Genetics. 2015;32(1):95-101.
51. Gilad Y, Bustamante CD, Lancet D, Paabo S. Natural selection on the olfactory receptor gene family in humans and chimpanzees. American Journal of Human Genetics. 2003;73(3):489-501.
52. Wu YL, Duan HP, Tian XC, Xu CS, Wang WJ, Jiang WJ, et al. Genetics of Obesity Traits: A Bivariate Genome-Wide Association Analysis. Frontiers in Genetics. 2018;9.
53. Bailey JNC, Palmer ND, Ng MCY, Bonomo JA, Hicks PJ, Hester JM, et al. Analysis of coding variants identified from exome sequencing resources for association with diabetic and non-diabetic nephropathy in African Americans. Human Genetics. 2014;133(6):769-79.
54. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol. 1994;235(2):625-34.
55. Cohen FE, Sternberg MJ. On the prediction of protein structure: The significance of the root-mean-square deviation. J Mol Biol. 1980;138(2):321-33.
56. Paul Campitelli TM, Sudhir Kumar, S. Banu Ozkan. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. **Annual Review of Biophysics**; 2020. p. 267-88
57. Liu F, Fitzgerald MC. Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using Limited Proteolysis. J Proteome Res. 2016;15(12):4666-74.
58. Smith IN, Thacker S, Seyfi M, Cheng F, Eng C. Conformational Dynamics and Allosteric Regulation Landscapes of Germline PTEN Mutations Associated with Autism Compared to Those Associated with Cancer. Am J Hum Genet. 2019;104(5):861-78.
59. Karn R, Emerson IA. Breast cancer mutation in GATA3 zinc finger 1 induces conformational changes leading to the closer binding of ZnFn2 with a wrapping architecture. J Biomol Struct Dyn. 2020;38(6):1810-21.
60. Xiao YM, Shaw GS, Konermann L. Calcium-Mediated Control of S100 Proteins: Allosteric Communication via an Agitator/Signal Blocking Mechanism (vol 139, pg 11460, 2017). Journal of the American Chemical Society. 2018;140(28):8998-.
61. Rose GD, Wolfenden R. Hydrogen bonding, hydrophobicity, packing, and protein folding. Annu Rev Biophys Biomol Struct. 1993;22:381-415.
62. Kuzmanic A, Zagrovic B. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. Biophys J. 2010;98(5):861-71.
63. Cheng X. Molecular Dynamics. In: Ivanov I, editor.: Methods in molecular biology; 2012. p. 243-85.

64. Bouard C, Terreux R, Tissier A, Jacqueroud L, Vigneron A, Ansieau S, et al. Destabilization of the TWIST1/E12 complex dimerization following the R154P point-mutation of TWIST1: an in silico approach. Bmc Structural Biology. 2017;17.
65. Khan MT, Rehaman AU, Junaid M, Malik SI, Wei DQ. Insight into novel clinical mutants of RpsA-S324F, E325K, and G341R of Mycobacterium tuberculosis associated with pyrazinamide resistance. Computational and Structural Biotechnology Journal. 2018;16:379-87.
66. Yoon JH, Nam JS, Kim KJ, Ro YT. Characterization of pncA mutations in pyrazinamide-resistant Mycobacterium tuberculosis isolates from Korea and analysis of the correlation between the mutations and pyrazinamidase activity. World Journal of Microbiology & Biotechnology. 2014;30(11):2821-8.
67. Sanjeev A, Mattaparthi VSK. Computational investigation on the effects of H50Q and G51D mutations on the α-Synuclein aggregation propensity. J Biomol Struct Dyn. 2018;36(9):2224-36.