Computational Cancer Genomics from a Systems-Biologic Perspective: From Sequence and Function to Pathways and Networks

Jimmy Lin[1], Christine Gan[1], Xiaosong Zhang[1], Siân Jones[1], Tobias Sjöblom[1], Laura D. Wood[1],
D. Williams Parsons[1], Giovanni Parmigiani[2], Nickolas Papadopoulos[1],
Kenneth W. Kinzler[1], Bert Vogelstein[1], Victor E. Velculescu[1]

Cancer is caused by sequential pathologic alterations in the genetic landscape. To understand this process, we recently sequenced over 13,000 genes in a panel of breast and colorectal cancers (Sjöblom et al. 2006). We found 1,149 genes with somatic mutations and described 191 candidate cancer genes. Here, we present the comprehensive analysis of these cancer genomes that goes beyond gene identification and explores systems-level characteristics. In total, we identified 35 sequence similarity clusters, 23 protein domains, 6 functional groups, 3 protein-protein networks, 12 interactome hubs, and 32 pathways, that are enriched for somatic mutations and are most likely to play a role in carcinogenesis and cancer progression.

From sequence and function to pathways and networks, we implemented numerous methods, tools, algorithms, and databases. For sequence, we performed pairwise BLAST for all genes and created networks of sequence similarity to identify highly enriched clusters. For protein domains, we used INTERPRO and Pfam annotation to look for protein domains that were preferentially mutated. For functional groups, we calculated over-representation of molecular function and biological processes from Gene Ontology. For protein-protein networks, we annotated interolog relationships in model organisms from HomoMINT and Ophid and constructed cancer interactomes. For hub calculation, we looked for genes that interacted preferentially to other mutated genes. For pathways, we examined the mutational frequency in defined pathways from KEGG, iPath, BioCyc, and BioCarta.

For each category, we used a combination of three different statistical models. We determined groups enriched in mutated genes by fitting the proportion of mutated genes in the study to the background dataset to a hypergeometric distribution. We calculated the enrichment in terms of mutations, by fitting the total numbers of base pairs sequenced for the group with numbers of mutations observed to a binomial distribution. This score not only takes into consideration gene number but also the cancer mutational prevalence. Lastly, we looked for profile enrichment by using the binomial probability of observing more than observed mutations in each gene and calculated the difference of the mutational profile for the gene group from background using gene set enrichment analysis (GSEA). The combination of these three methods allows the identification of groups that are significant in terms of number of genes mutated, number of mutations observed, and the mutational profile of the group.

This comprehensive systems exploration identified significant members for sequence similarity clusters, protein domains, functional groups, protein-protein networks, interactome hubs, and pathways. The summation of all the different analyses provide a multi-perspectival framework to guide further research and help identify cellular processes critical for malignant progression as well as therapeutic intervention.

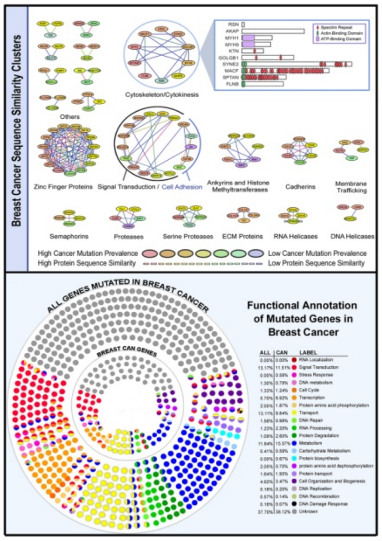
Breast Cancer Sequence Similarity Clusters
RSN
AKAP
MYH1
MYH9
KTN
GOLGB1
SYNE2
MACF
SPTAN
FLNB
Spectrin Repeat
Actin-Binding Domain
ATP-Binding Domain
Cytoskeleton/Cytokinesis
Others
Zinc Finger Proteins
Signal Transduction / Cell Adhesion
Ankyrins and Histone Methyltransferases
Cadherins
Membrane Trafficking
Semaphorins
Proteases
Serine Proteases
ECM Proteins
RNA Helicases
DNA Helicases
High Cancer Mutation Prevalence
Low Cancer Mutation Prevalence
High Protein Sequence Similarity
Low Protein Sequence Similarity
ALL GENES MUTATED IN BREAST CANCER
BREAST CAN GENES
Functional Annotation of Mutated Genes in Breast Cancer
ALL | CAN | LABEL
0.00% | 0.00% | RNA Localization
13.17% | 11.51% | Signal Transduction
0.00% | 0.58% | Stress Response
1.30% | 0.79% | DNA metabolism
1.33% | 1.24% | Cell Cycle
5.70% | 6.92% | Transcription
2.05% | 1.87% | Protein amino acid phosphorylation
13.11% | 9.84% | Transport
1.56% | 0.98% | DNA Repair
1.23% | 2.20% | RNA Processing
1.09% | 2.60% | Protein Degradation
11.84% | 15.57% | Metabolism
0.41% | 0.59% | Carbohydrate Metabolism
0.00% | 0.67% | Protein biosynthesis
2.05% | 0.70% | protein amino acid dephosphorylation
1.64% | 1.95% | Protein transport
4.92% | 3.47% | Cell Organization and Biogenesis
0.16% | 0.20% | DNA Replication
0.57% | 0.14% | DNA Recombination
0.16% | 0.07% | DNA Damage Response
37.70% | 38.12% | Unknown

Figure 1. Sequence Similarity Clusters and Gene Ontology Annotation of Genes Mutated in Breast Cancer

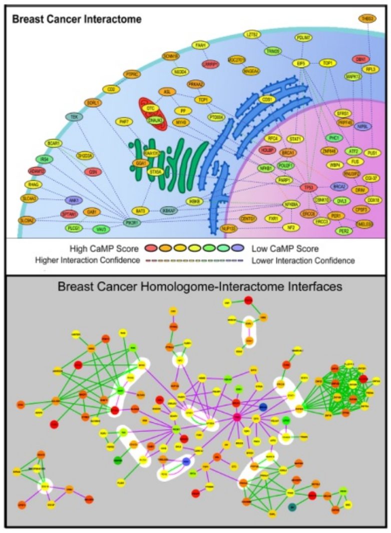

Figure 2. Protein-Protein Interaction Network and Intersection with Sequence Similarity Clusters of the Breast Cancer Genome

REFERENCE: Sjöblom, T., S. Jones, L.D. Wood, D.W. Parsons, J. Lin, T.D. Barber, D. Mandelker, R.J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S.D. Markowitz, J. Willis, D. Dawson, J.K. Willson, A.F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B.H. Park, K.E. Bachman, N. Papadopoulos, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314: 268-274.

[1]Ludwig Center for Cancer Genetics and Therapeutics, and The Howard Hughes Medical Institute at The Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA.
[2]Departments of Oncology, Biostatistics and Pathology, The Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA.