A novel computational method for finding regions with copy number abnormalities in cancer cells

Vivek, Manuel Garber, and Mike Zody

Broad Institute of MIT and Harvard, Cambridge, MA, USA

Introduction

Cancer can result from the over expression of oncogenes, genes which control and regulate cell growth. Sometimes oncogenes increase in activity due to a specific genetic mutation called a translocation (**Fig 1**). This translocation allows the oncogene to remain as active as its paired gene. Amplification of this mutation can occur, thereby creating the proper conditions for uncontrolled cell growth; consequently, each component of the translocation will amplify in similar quantities. In this mutation, the chromosomal region containing the oncogene displaces to a region on another chromosome containing a gene that is expressed frequently.

Fig 1. Two chromosomal regions (abcdef and ghijk) are translocating to create two new regions (abckl and ghijedf).

Fig 2. FISH sample mapping

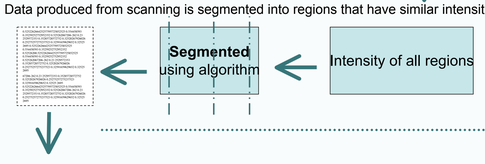
Traditional methods, such as FISH, present scientists with a visual representation of chromosome arrangements (**Fig 2**). Newer Array Hybridization experiments (e.g. Affymetrix SNP arrays, Nimblegen isothermal oligo arrays, BAC Arrays) provide relatively cheap, easy and increasingly high resolution methods to detect a change in copy number of a sample when compared to appropriate controls. However, how these duplications are arranged in the sample genome remains undetectable with these newer methods. Raw data produced from Affymetrix SNP arrays can provide clues as to how genes are amplified. These data sets give the intensities for each chromosomal region in each sample. We believe that those regions possessing similar amplification levels may have amplified together in a translocation and those with similar deletion levels were knocked out in order to promote cell growth. Thus, we created a computer program that determines which regions of the genome correlate by copy number in cancer cells in order to detect translocations.

Methods

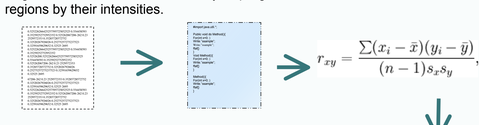
A dataset consisting of 129 colorectal cancer samples and 6319 segments was used to produce results.

Samples are extracted from patient and scanned using Affymetrix scanners.

Data produced from scanning is segmented into regions that have similar intensities



Segmented data is fed into an analysis program which correlates the chromosomal regions by their intensities.



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

Correlations are given scores, and highly correlated regions are checked against the UCSC Genome Browser.

1 ~0.49 2 ~0.47 3 ~0.41

Fig 1. Chromosome Mutations WebQuest. *Carmel High School Biological Sciences*

Fig 2. Array-based comparative genomic hybridization and copy number variation in cancer research. *Cytogenet Genome Res. 2006;115(3-4):262-72.*

Gene Function snippets come from the UCSC Genome Browser.

Results

1 SMAD4 – a gene known to be deleted in pancreatic carcinoma

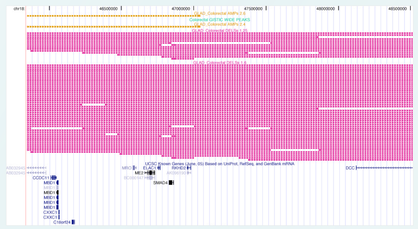
Results from Analysis Program

Region 1	Region 2	R^2
Chr18:47044749-47311978	Chr17:13930739-14654741	0.499070821478475

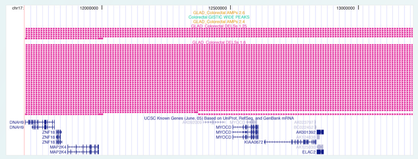
Actual region containing gene

chr18: 45,842,214 - 48,514,513

Genes on Chr18:45,842,214-48,514,513[1] (UCSC Genome Browser Snapshot)



Genes on Chr17: 11,654,742 - 13,500,000[1] (UCSC Genome Browser Snapshot)



➢**AK001392** - prostate cancer protein (**see gene ranking 3**)

- Regions known to be **amplified** based on prior data
- Regions known to show high **amplifications** or **deletions**
- Regions known to be **deleted** based on prior data

2 COX10 – a gene deleted in cytochrome c oxidase deficiency, known to be related to cell proliferation

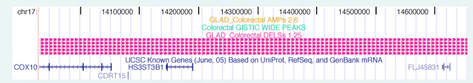
Results from Analysis Program

Region 1	Region 2	R^2
Chr17:13930739-14654741	Chr18:26861790-27072166	0.47355172850856

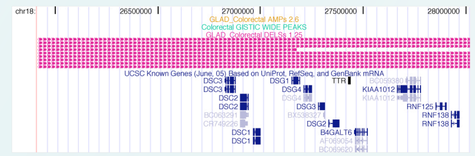
Actual region containing gene

chr17: 13,966,862 - 14,068,461

Genes on Chr17: 13,930,739 – 14,654,741 (UCSC Genome Browser Snapshot)



Genes on Chr18: 25,915,093 - 28,018,862[1] (UCSC Genome Browser Snapshot)



➢**DSCs** – "calcium-dependent glycoprotein" in epithelial cells
➢**DSGs** – proteins found in "cell-cell junctions between epithelial, myocardial and certain other cell types "
➢**KIAA1012** – "may play a role in vesicular transport from endoplasmic reticulum to Golgi"
➢**RNFs** – ring finger proteins "known to be involved in protein-DNA and protein-protein interactions"
➢**B4GALT6** – "one of seven beta-1,4-galactosyltransferase" genes, encodes an enzyme important for glycolipid photosynthesis

[1] Denotes either an adjacent region or a zoomed out view of the region. This region contains genes that may have some relevance to cancer. Adjusting the view within the Genome Browser was necessary to detect these genes.

3 AK001392 – a hereditary prostate cancer protein

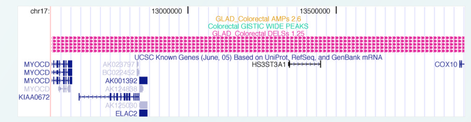
Results from Analysis Program

Region 1	Region 2	R^2
Chr17:12542326-13930738	Chr8:1789292-1801984	0.406208680312004

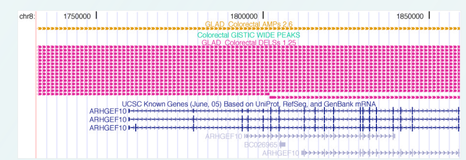
Actual region containing gene

chr17: 12,542,326 - 13,930,738

Genes on Chr17: 12,542,326 – 13,930,738 (UCSC Genome Browser Snapshot)

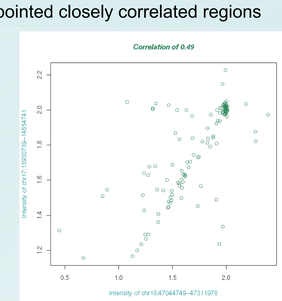


Genes on Chr8: 1,789,292 – 1,801,984 (UCSC Genome Browser Snapshot)



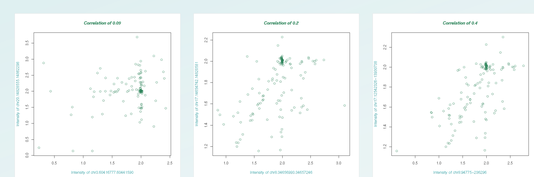
➢**ARHGEF10** – "play[s] a fundamental role in numerous cellular processes that are initiated by extracellular stimuli that work through G protein coupled receptors"

➢A correlation matrix of the region intensities accurately pinpointed closely correlated regions



Regions on chromosomes 17 and 18 are shown to have a high correlation of ~0.49. In this specific correlation, the two regions are shown to be deleted. SMAD4 is a known cancer related gene depicted in the chromosome 18 region. (See SMAD 4 section)

➢Correlations were found to be highly distinguishable in activity



R^2 ~0.09 R^2 ~0.2 R^2 ~0.4

Left is a graph showing correlation of 0.09. Center is a graph showing correlation of 0.2. Right is a graph showing correlation of 0.4.

Acknowledgements

I'd like to thank my mentors Manuel Garber and Mike Zody, as well as our internship directors Megan Rokop, Julie Boehm, and Kate MacSwain for making this internship experience possible. I'd also like to thank Melissa Parkin for showing me how the data was produced with the Affymetrix scanners and Eva Otero and Jessica Perez for giving me an introduction to the mechanical engineering side of the Broad.

Conclusion and Future Directions

It is possible to use data produced from Array Hybridization experiments to make predictions regarding chromosome translocations. Using correlation, we figured that the deleted regions with high copy number correlation were necessary deletions for cancer to thrive. We checked whether some of these genes were possible cancer genes (tumor supressors/oncogenes) with gene descriptions found on the UCSC Genome Browser and found that they were known to play a role in cancer as genes known to be deleted for cell proliferation to take place. Unfortunately, we could not find regions that had amplified, and thus, could not hypothesize any translocations. Though, it is important to note that we discovered high correlations where both regions were deleted together. It is interesting that SMAD4, a pancreatic carcinoma related gene, and COX10, a mitochondrial gene, were highly ranked in our colorectal cancer data. According to prior research, defects in SMAD4 are known to be associated with an increased risk of colon and gastrointestinal cancers. Furthermore, the mitochondria plays a role in cellular apoptosis. This function of the mitochondria would suggest that defective mitochondria would be unable to prohibit cell proliferation.

➢Perhaps there exists links between pancreatic carcinomas, mitochondrial deficiencies and colorectal cancer that should be further researched.

➢Furthermore, the fact that both regions and not just one region in the correlations are being deleted begs for a biological explanation. Does a double hit (deletion of partners of SMAD4 and COX10) increase cancer fitness more than the deletion of each of them alone?

➢In addition, although this project aimed to determine the existence of a translocation, we are still in the process of statistically testing the significance of these results.

➢Lastly, the results were derived solely from computational correlation. Pure biological means will be needed to verify any amplification results. Thus, we are looking for a partner to test these high correlated regions through PCR of these significant regions.