Thesis for the degree
Doctor of Philosophy

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
דוקטור לפילוסופיה

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Yotam Drier

מאת
יותם דרייאר

ניתוח סידור מחדש של הדנ"א, והערכת מידת השיבוש של תהליכים ביולוגיים בסרטן

Analyzing somatic DNA rearrangements, and quantifying pathway deregulation in cancer

Advisor:
Prof. Eytan Domany

מנחה:
פרופ' איתן דומאני

November 2012

כסלו התשע"ג

Table of Contents

1. Acknowledgements

It is a pleasure to thank my supervisor Prof. Eytan Domany. I have benefited enormously from his devoted guidance, knowledge and insight. His personality and approach have made these years a wonderful and rewarding experience. I would also like to thank Dr. Gaddy Getz, who has been an unofficial, yet extremely significant mentor. Gaddy's support and guidance had allowed me to understand how "big science" is being done, and how one can harvest the benefits of taking part in big projects in the frontier of biomedical research to perform creative and independent research. In addition, I owe a debt of gratitude to Eytan's support and Gaddy's hospitality, which not only shaped my professional life, but also allowed me to meet and be with my wife despite the geographical challenge.

I would also like to thank all those that I had the fortune to collaborate with in my work, and especially Dr. Yaara Zwang and Sagi Nahum. Special thanks to Dr. Amos Tanay, Dr. Rotem Sorek and Prof. David Mukamel for their helpful advice and guidance throughout my work.

I would like to thank my colleagues at the Broad Institute for their help and hospitality, including Prof. Matthew Meyerson, Prof. Levi Garraway, Prof. Rameen Beroukhim, Prof. Eric Lander, Mike Lawrence, Mike Berger, Cheng-Zhong Zhang, Chip Stewart, Sylvan Baca, Adam Bass, Scott Carter, Kristian Cibulskis, Douglas Voet and Petar Stojanov. In addition I would like to thank past and present members of the Domany group, for their professional advice as well as moral support, including Michal Sheffer, Noa Bossel, Amit Zeisel, Assif Yitzhaky, Tal Shay, Yuval Tabach, Rita Vesterman, Barak Markus, Libi Hertzberg, Ofer Tabach, Asaf Farhi, Lior Haviv Gelibter, Anna Llivshits, Yair Horesh, Noam Shental, Hilah Gal, Michael Bon and Roni Golan-Lavi.

My deep gratitude goes to my wife Michal, my parents and the rest of my family and friends for caring, encouraging, and making it all worth the while.

2. List of abbreviations

mRNA – Messenger ribonucleic acid

DNA - Deoxyribonucleic acid

WGS - Whole genome sequencing

NHEJ - Non-homologous end joining

MMEJ - Microhomology-mediated end joining

MAQ - Mapping and assembly with quality (alignment software)

BWA - Burrows-Wheeler aligner (alignment software)

IGV - Integrative genomics viewer (visualization software)

GBM - Glioblastoma multiforme

FDR - False discovery rate

PCA - Principal component analysis

ANOVA - Analysis of variance

MAD - Median absolute deviations

REMBRANDT - Repository for molecular brain neoplasia data

TCGA - The cancer genome atlas

NCI - National cancer institute

KEGG - Kyoto encyclopedia of genes and genomes

PID - Pathway interaction database

CIN - Chromosomal instability

MSI - Microsatellite instable

MSS - Microsatellite stable

MV-CAD – Multi vessel coronary artery disease

NC - Normal coronaries

MI - Myocardial infarction

STEMI - ST Segment elevation myocardial infarction

PBMC - Peripheral blood mononuclear cells

3. Abstract

In my doctoral studies I have focused on genome-wide study of cancer using computational tools and statistical analysis. My motivation was interesting biological questions that require comprehensive and complex analysis, as I found this the best way I could do interesting and challenging science, which may also have high impact and contribution. Moreover, I have specific interest in the study of cancer, as I feel that there is still much more to learn from a comprehensive systematic genome wide study of cancer, and that it may have a high impact on human health and wellbeing. I have applied this philosophy in two major research projects: The first is to develop a method to pinpoint rearrangement breakpoints in cancer, and to apply it to study somatic rearrangements, as described in chapters 4 and 5. The second is to develop a method to quantify pathway deregulation, and apply it to study pathway deregulation in glioblastoma (aggressive brain cancer), as described in chapter 6. Furthermore, I have been involved in the design and analysis of several smaller scale experiments, two of them described in chapters 7 and 8.

תקציר

בלימודי הדוקטורט שלי התמקדתי בחקר הגנום של סרטן באמצעים חישוביים וניתוח סטטיסטי. המוטיבציה שלי נבעה מחיפוש אחר שאלות ביולוגיות מעניינות שדרושות ניתוח מעמיק ומורכב, היות שמצאתי שזו הדרך הטובה ביותר לעסוק במחקר שגם מאתגר ומרתק אותי וגם תורם תרומה משמעותית לקידום המדע. יתר על כן, מצאתי עניין מיוחד בחקר הסרטן, משום שאני מאמין שבתחום זה נדרשת עבודה אנליטית מורכבת שיכולה לפרוץ דרך בהבנתנו את הסרטן, וכן מכיוון שאני חש שבתחום זה ניתן לקדם גם את המדע הבסיסי וגם את הרפואה באופן שישפיע משמעותית על בריאות האדם. דרך חשיבה זו הנחתה אותי בשני הפרויקטים המחקרים העיקריים שעסקתי בהם- הראשון לפיתוח שיטה למציאת נקודות השבירה בדנ"א הסרטני וליישומה לחקר המאפיינים והמנגנונים של סידור מחדש של הדנ"א בסרטן, כפי שמתואר בפרקים 4 ו 5; בשני פיתחתי שיטה להערכת מידת השיבוש של תהליכים ביולוגיים בסרטן ויישומה לחקר גליובלסטומה (סרטן מח אלים), כפי שמתואר בפרק 6. בנוסף, הייתי מעורב בתכנון וניתוח מספר ניסויים בקטנה מידה קטן יותר, שניים מהם מתוארים בפרקים 7 ו 8.

4. Detecting somatic rearrangements across cancer reveals classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability

In collaboration with members of the Cancer Program at the Broad Institute of MIT and Harvard. Publication #1.

Introduction

Alterations in DNA drive much of cancer development. Many of these alterations are "structural", leading to fusions between distant regions of the genome. Many alterations are deletions and amplifications, which introduce copy-number changes. Others, such as inversions and balanced translocations, maintain copy number. Multiple mechanisms can cause these alterations, including deterioration of DNA repair and replication mechanisms[1, 2]. A rearrangement can be thought of as two (related) processes – one is the breakage of the DNA, and the other is fusing the DNA together, but failing to restore the original DNA structure. Each rearrangement, therefore, defines two DNA *breakpoints*.

Recently, whole genome sequencing (WGS) became affordable enough to allow mapping of rearrangements for large cancer cohorts. This provides the opportunity to answer several key questions on DNA breakage in cancer. We and others have started to approach this by analyzing samples from specific tumor types[3-13]. Whole genome sequencing was done by paired end sequencing using Illumina HiSeq. DNA samples are taken from the tumor and normal blood from each patient, and are fragmented to a *library* of 500-700bp long DNA *fragments*. Each fragment is sequenced from both ends, yielding a pair of 101bp *reads*, leaving an unknown sequence in the middle (whose size is referred to as *insertion length*). The sequencing data is then preprocessed by the Picard and Firehose pipelines[7], including mapping of the read pairs to the reference genome by an off the shelf alignment algorithm (MAQ or BWA). The rearrangement pipeline is composed of dRanger – a method we developed to predict rearrangement regions, and BreakPointer – a method I developed to pinpoint the rearrangement breakpoints (see Methods). This was used for functional

analysis of several types of cancer (see chapter 5), and for the study described here, analyzing breakpoint patterns across cancer (95 samples of 7 types of cancer).

I found three genomic factors that significantly affect the distribution of DNA breakpoints across cancer along the genome: replication time, proximity to transcribed genes and the GC content. These correlations allow us to hypothesize about the causes and cell-cycle timing (mitosis / interphase) of the breakage events, and serve as a basis for future modeling of passenger rearrangements in cancer (only a small fraction of the rearrangements are drivers of the malignant transformation). I also identified a significant correlation between breakpoints and somatic point mutations. Although formally I cannot distinguish between cause and effect, I ruled out the possibility that the correlation is merely due to genomic variation in the susceptibility to acquire both types of genome alteration. Furthermore, pinpointing the precise breakpoints of rearrangements allows characterization of microhomology (see Figure 4-1 below), which may suggest potential mechanisms of rearrangement.

This work is summarized in a manuscript accepted to Genome Research, which appears as publication 1 of the list presented in Chapter 9.

Results

Detecting somatic rearrangements

The growing number of whole genome sequencing efforts in cancer is raising the need to accurately pinpoint rearrangement breakpoints without additional experimental measurements, particularly due to the high number of breakpoints found. Several studies to date[3-6] either published approximate breakpoint locations, or performed additional experiments to pinpoint the breakpoints (e.g. by amplification of the region and resequencing). We recently published several other studies[7-13] (see also chapter 5) in which we pinpoint the breakpoints to base-pair resolution using BreakPointer, described here in detail (Methods, Figure 4-11).

In this study, I perform a pan-cancer analysis of rearrangement breakpoints based on WGS data from 95 matched tumor/normal samples: 24 breast samples sequenced at the Sanger Institute[4] and 71 sequenced at the Broad Institute from various tumor types: 23 multiple myeloma[8], 22 breast carcinomas[13] , 9 colorectal carcinomas[9] , 7 prostate[7], 5 melanoma[12], 3 chronic lymphocytic leukemia[10], and 2 head and neck[11]. A total of 4996 candidate and approximate somatic rearrangements were detected using

dRanger (Methods) in the 71 Broad Institute samples. Out of these, 4368 (87%) were successfully pinpointed to single base pair resolution using BreakPointer. I successfully validated the existence of 1580 out of 1880 (84%) rearrangements randomly selected for validation, by PCR and targeted pyrosequencing (Methods), and confirmed the exact pinpointing of 1503 (95%) of them by aligning the pyrosequencing results to the fused sequence predicted by BreakPointer. In the analyses presented below, I use different datasets -- either the 4368 successfully pinpointed breakpoints or, when relevant, the 4996 candidate rearrangements (though the additional 628 rearrangements do not significantly change the results). The additional 24 samples by Stephens et al.[4] are used only for the analysis of factors determining the distribution of breakpoints.

Microhomology of rearrangements

Rearranged DNA segments occasionally share a short stretch of identical sequence, known as an overlapping microhomology[14], see Figure 4-1. The base pairing between the two segments being fused guides the exact location of the fusion. Knowing the exact breakpoint allows measuring the microhomology for every rearrangement.

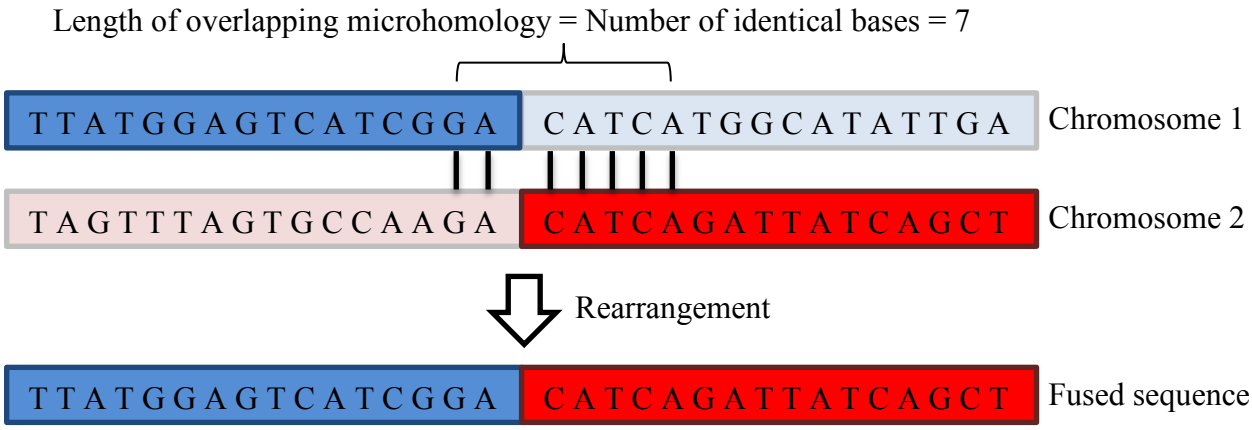


Figure 4-1 – **An example of a translocation between the blue and red regions depicts the definition of overlapping microhomology, shown here by black lines**. The two sequences above the arrow show the two regions before the rearrangement, and on the bottom the resulted fused sequence is shown.

In our cohort, somatic rearrangements display an increased level of microhomology, with an average of 1.7bp instead of the 0.7bp expected by chance (a 2.4-fold increase, Wilcoxon p-value $<10^{-250}$, see Methods). To study whether this excess of homology occurs in all types of rearrangements, I classified them into five categories: (i) Short

deletions (<5Kb); (ii) Inversions; (iii) Tandem duplications; (iv) All other intra-chromosomal rearrangements (mostly deletions); and (v) Inter-chromosomal translocations. All types showed more microhomology than expected by chance (2.2 – 2.8 fold increase, Wilcoxon p-values<10^{-25}). This is true also for every type of cancer separately- except intra-chromosomal rearrangements in CLL, all types with 10 or more rearrangements showed significant increase, FDR<10%. The short microhomologies imply the involvement of non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ) in almost all somatic rearrangements (only 0.2% of detected rearrangements displayed more than 20bp homology). MMEJ is rare, while NHEJ is quite frequent (only 2.5% of rearrangement had more than 5bp microhomology, 44.2% at least 2bp but at most 5bp).

Even when comparing only to non-homologous germline rearrangements in 185 human genomes[15], I find that the microhomologies of somatic rearrangements were shorter (average of 1.7bp vs. 2.2bp, Mann–Whitney p-value<5.4×10^{-14}), and MMEJ less frequent (6.6% of non-homologous germline rearrangements had more than 5bp microhomology, 46.6% at least 2bp but at most 5bp). Recently complex rearrangements in the germline were characterized in several individuals[16], which showed less microhomology than Mills et al. These complex germline events are closer to the somatic events described here in terms of the overall microhomology distribution (average 1.43bp, Mann–Whitney p-value<0.012), probably due to less NHEJ and more MMEJ (5.7% had more than 5bp microhomology, 28.6% at least 2bp but at most 5bp).

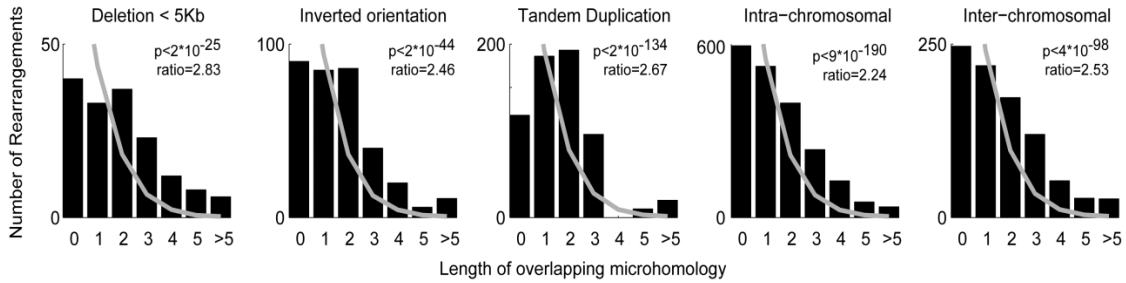


Figure 4-2 - **Overlapping microhomology by rearrangement type.** Gray line represents the expected distribution, by permuting rearrangement pairs. All rearrangement types show higher microhomology than expected by chance. Tandem duplications display the highest microhomology rate, and in fact microhomology of length 2 is the most common case. Short deletions (up to 5K) and inversions show more microhomology than other rearrangements. Scholz-Stephens p-value for significant difference between histograms is < 10^{-6}.

The distribution of microhomology lengths varied by the type of rearrangement (Scholz-Stephens' p-value<10^{-6}, see Methods). Tandem duplications had the most distinctive distribution, with 2bp (typical for non-homologous end joining) being the most common overlap across all tumor types (as we previously reported in colorectal cancer[9]). Short deletions and inversions displayed a similar pattern (Figure 4-2). Differences in microhomology, and specifically more frequent microhomologies in tandem duplications was previously reported for breast cancer[4].

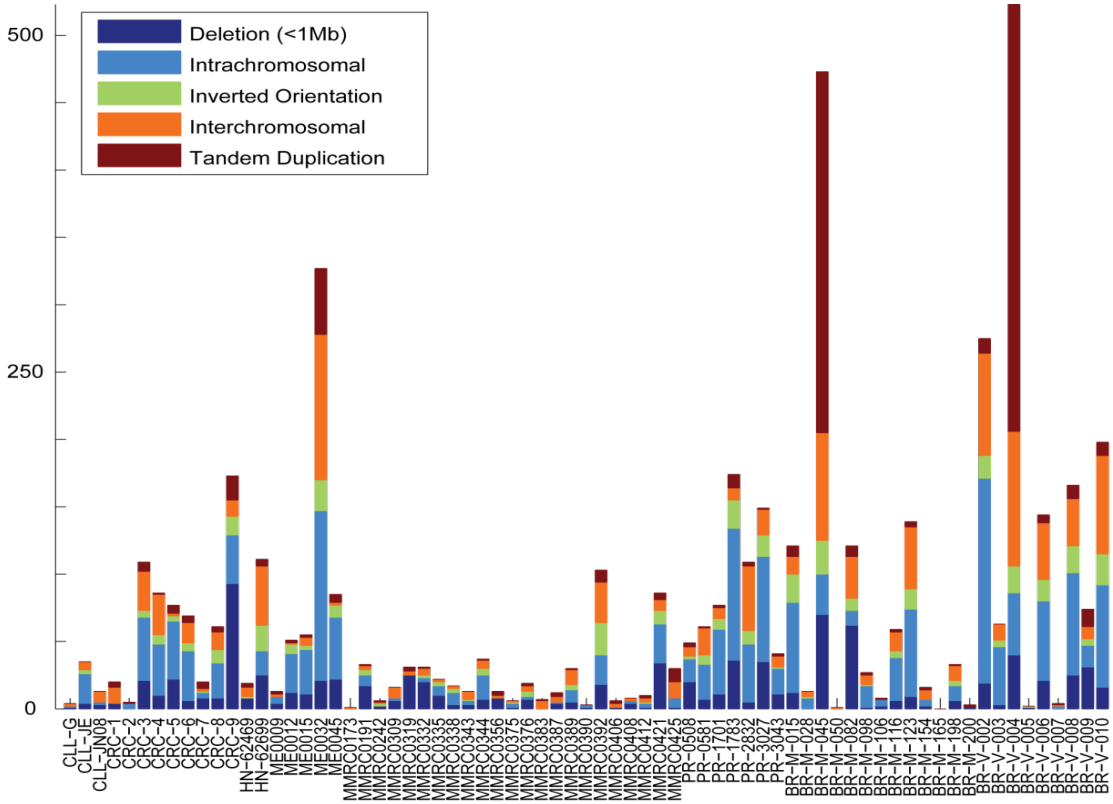


Figure 4-3 – **Different types of rearrangements across 71 samples with a total of 4996 rearrangements.** Interesting outliers are CRC-9 with many short deletions and a few inter-chromosomal rearrangements, and on the other end of the spectrum-ME0032 with a few short deletions and many inter-chromosomal rearrangements. Also noticeable are the large number of tandem duplications found in BR-M-045 and BR-V-004.

Each sample had a different composition of rearrangement types (Figure 4-3), and therefore differences between the microhomology distributions of different samples are to be expected. However, even when controlling for the sample-specific composition and using the overall microhomology pattern for each type, 6 of the 71 samples (8%) still had a significantly different distribution (FDR<4%, Figure 4-4, Methods). Three prostate samples displayed less microhomology than expected by

their composition, while three breast samples displayed more, suggesting mechanistic differences not only between the different types of rearrangements but also between prostate, breast and other cancers. Indeed when pooling all breast samples together they show more microhomology than expected by their composition ($p<10^{-6}$), and all prostate samples pooled show less microhomology than expected ($p<10^{-6}$).

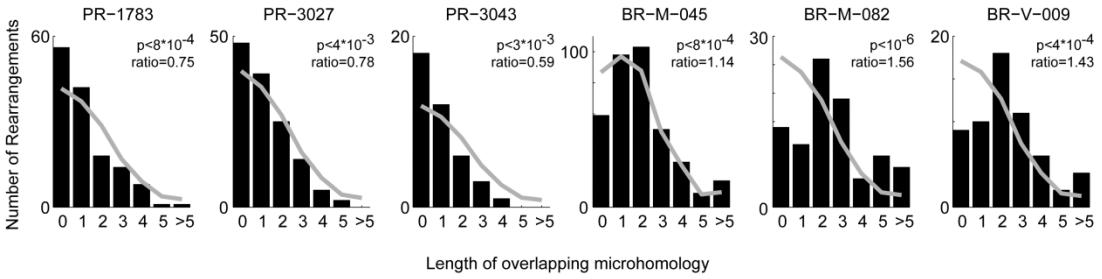


Figure 4-4 - **Overlapping microhomology of six extreme samples.** Gray line represents the expected distribution, by the composition of the different rearrangement types. The three prostate samples show less microhomology than expected (notice the high fraction of breakpoints with no microhomology), and the three breast samples shows more (low fraction of breakpoints with no microhomology). Expected distribution was constructed to control for the different rearrangement types and the homologies they display. These are the only samples passing FDR<10% (and in fact satisfy FDR<4%).

Factors determining the distribution of breakpoints

Next, I examined genomic features to identify ones that may affect, or at least, are correlated with the density of rearrangement breakpoints along the genome. First, I examined whether the distribution of breakpoints was correlated with local transcription levels typical for that tumor type (Methods). As for microhomolgies, strong sample-specific effects are observed, with different samples showing opposite behaviors -- some with significant enrichment of breakpoints near transcribed genes (most pronouncedly within 10Kbp-100Kbp windows) and others with significant depletion (Figure 4-5).

Subsequently, I examined the correlation with two additional factors that may affect the location of rearrangements – DNA replication time and GC content. I first considered the effect of each factor separately and partitioned the genome into three or four distinct parts according to the level of each factor. I then calculated, for each sample, the relative rate of breakpoints in every part of the genome (represented as log fold-change to the genome-wide average) and a significance level (Figure 4-6-A, Methods). Interestingly, the majority of samples showed enrichment of breakpoints

either at early replicating, high %GC, transcribed regions of the genome (EHT) or at late replicating, low %GC, untranscribed (LLU) regions. The fact that the effects of these three variables are correlated is not surprising since they are mostly correlated along the genome. Studying the enrichment patterns across cancer revealed tumor-type specific patterns -- CLL and breast cancer samples tend to have breakpoints at EHT regions, while colorectal cancer, melanoma and head and neck cancer samples tend to have breakpoints in LLU regions (Figure 4-6-B).

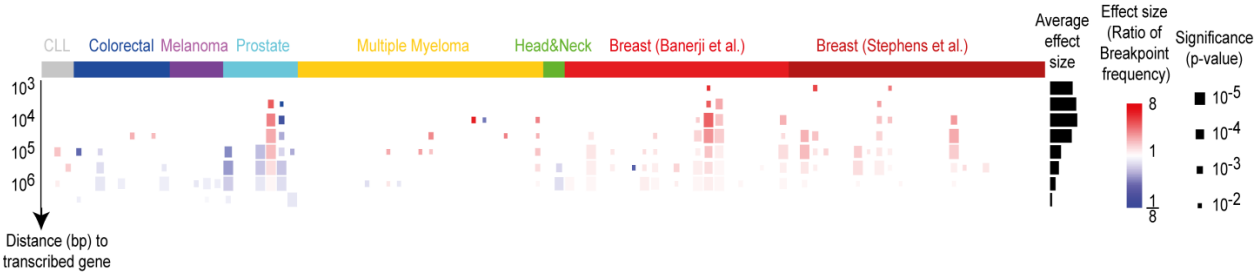


Figure 4-5 - **Breakpoints in transcribed and untranscribed regions.** Each square represent enrichment (red) or depletion (blue) of breakpoints in transcribed regions defined by maximal distance to transcribed gene. Size represents p-value, and color represent ratio. Only tests that passed 10% FDR are shown. Notice that regions of approximately 10^4-10^5 bp often are significantly enriched or depleted. On the right the average ratio (across samples) is shown. The colored bar above specifies the type of cancer for each sample.

Four samples showed contradictory patterns of LLU and EHT, deviating from the above pattern, suggests that, at least in these cases, more than one factor is required to explain the density of breakpoints. The colorectal sample CRC-3 was enriched for breakpoints in late-replicating, untranscribed regions, but depleted in regions of low %GC. The multiple myeloma sample MMRC0421 and melanoma sample ME0032 harbored breakpoints in untranscribed regions, but also in regions with high GC content, and the breast sample PD3668a was depleted in both low %GC and high %GC.

These inconsistent patterns can be somewhat explained by examining the contribution of each type of rearrangement separately. Surprisingly, in these samples, different types of rearrangements follow different patterns of enrichment. For the melanoma sample ME0032 (Figure 4-7) inter-chromosomal translocations and intra-chromosomal inversions and tandem duplications were enriched in regions of high %GC, while other intra-chromosomal events were skewed towards low %GC and untranscribed regions. Similarly, for multiple myeloma sample MMRC0421, intra-

chromosomal rearrangements contributed to enrichment in untranscribed, low %GC regions, while inversions were enriched in high %GC.

In order to quantify the joint contribution of all three parameters and attain a compact representation, I used logistic regression (Methods, Figure 4-8). This type of analysis requires a large number of rearrangements in order to uncover significant results, and so only the most highly rearranged samples are amenable. To cope with this challenge, I pooled together several samples of the same cancer type. In contrast to the outliers described above, it seems that in general rearrangements of different types (deletion, inversion, etc.) are distributed similarly to each other; however, possibly this is due to some cancellation of opposite effects caused by pooling of samples.

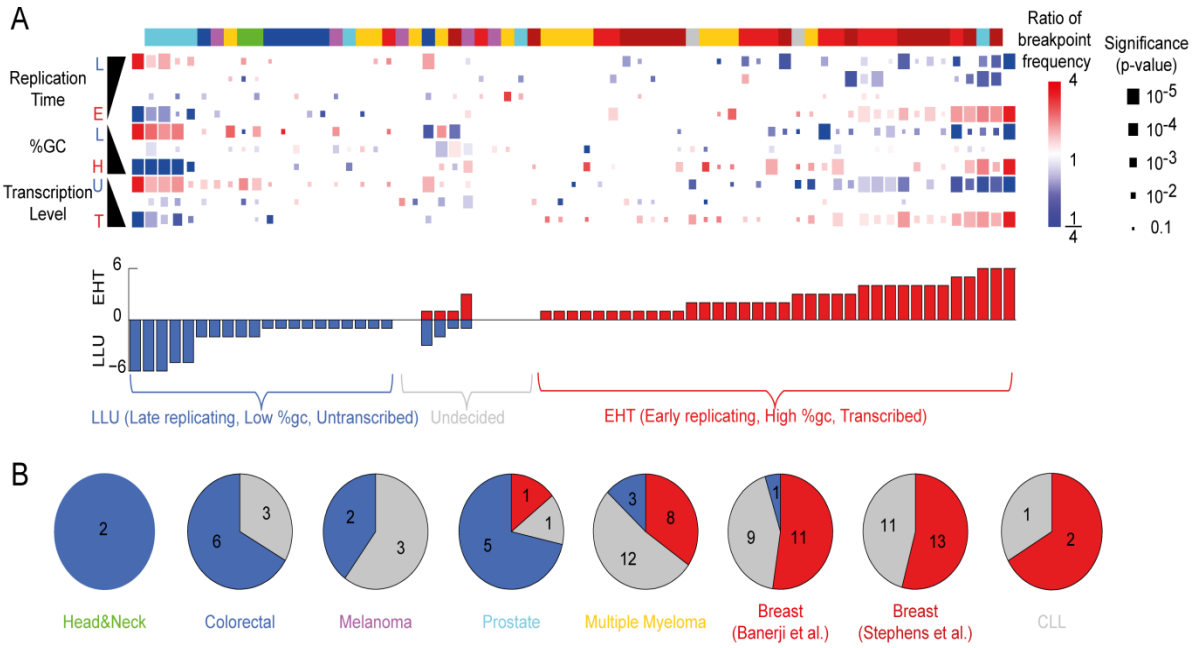


Figure 4-6 - **Breakpoint distribution as function of transcription, replication and GC content across samples. A.** Each row represent different bin of replication time, GC content or distance from transcribed gene. Each square represent significant (FDR<10%) enrichment or depletion, size represents p-value, and color represent ratio. Only samples with at least one significant bin are shown. The colored bar above specifies the type of cancer for each sample. Most samples are either enriched for breakpoints in early replicating, high %GC, transcribed regions of the genome (EHT) or in late replicating, low %GC, untranscribed regions (LLU), as can be seen in the bar chart. The samples are sorted by the agreement with that pattern. **B.** The breaking of each cancer to EHT (red) and LLU (blue) samples. Samples without any significant extreme bin, or with contradicting enrichments are shown in gray.

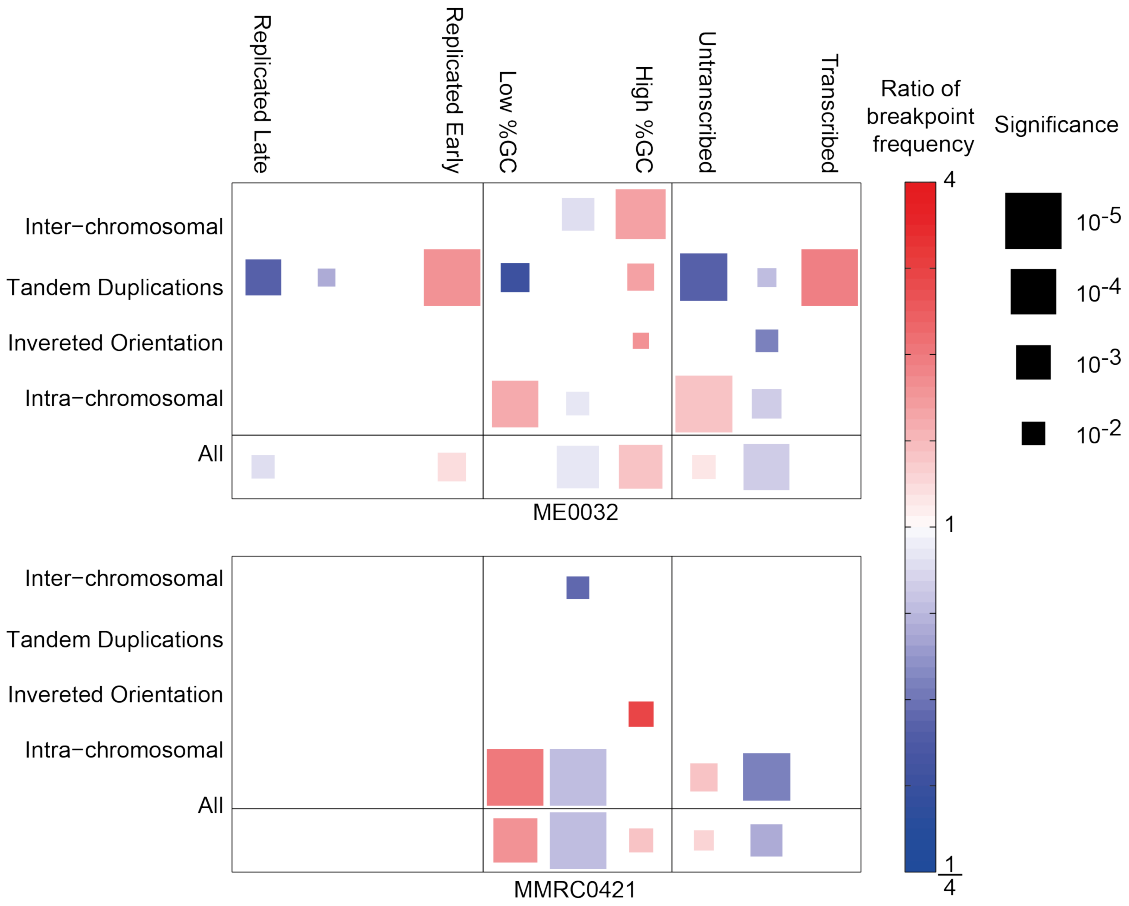


Figure 4-7 – **Analysis of genomic factors broken down by rearrangement type**. Enrichment and depletion of melanoma sample ME0032 and multiple myeloma sample MMRC0421. Each column represents a different bin of replication time, GC content or distance from active gene. Each square represents a significant (FDR<10%) enrichment or depletion, its size represents the p-value, and its color represents the enrichment ratio. Enrichment of breakpoints in high %GC and in untranscribed regions suggests that complex opposing forces are at play. Indeed for ME0032, Inter-chromosomal translocation, Tandem duplications, and inverted intra-chromosomal rearrangements tend to be in high %GC, while other intra-chromosomal rearrangements (mostly deletions) tend to be in untranscribed regions of the genome. For MMRC0421, Inverted intra-chromosomal rearrangements tend to be in high %GC, while other intra-chromosomal rearrangements (mostly deletions) tend to be in untranscribed (and low %GC) regions of the genome.

Next, I searched for genes with mutations that are correlated with the different patterns of rearrangements (LLU or EHT). Interestingly, I identified APC as the only gene whose mutations (in the coding region or promoter) are significantly associated with the LLU enriched samples ($q<0.05$; Methods). The adenomatous polyposis coli (*APC*) gene is mutated in 8 of the 71 samples, 7 of which are included the 19 LLU samples (Fisher's exact test; $p=10^{-4}$, $q=0.017$). *APC* binds to and stabilizes microtubules, and is necessary to keep chromosomal integrity during mitosis[17, 18]. Defects in APC might, therefore, lead to chromosome breakage during aberrant

mitosis, or disrupt mechanisms that protect or repair heterochromatin regions. APC is known to be highly mutated in colorectal cancers (~70-80%)[19, 20] and indeed, all 6 colorectal samples with an APC mutation were LLU and the remaining 3 were not. This explains the high prevalence of LLUs in colorectal cancer. This might suggest that the correlation to APC mutations is merely due to colorectal cancer being a confounding variable, and require further study on larger cohorts.

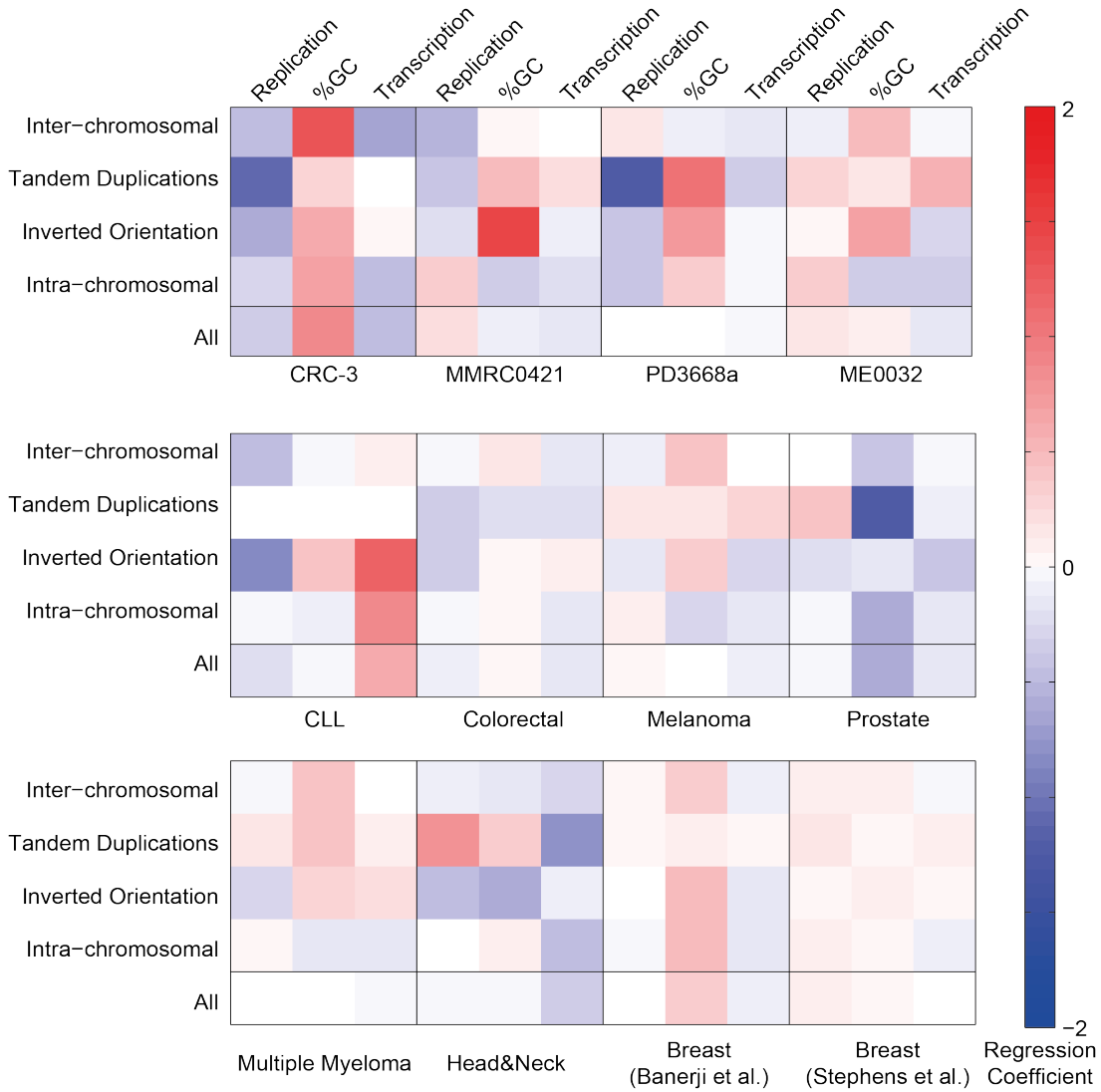


Figure 4-8 – **Logistic regression of the three genomic factors broken by rearrangement type.** Positive coefficient (red) means that the feature contributes to early replicating, high %GC, transcribed regions of the genome, while negative coefficient (blue) stands for depletion. The top row displays the four samples with contradicting enrichments, showing that while CRC-3 displays similar behavior for all types of rearrangements, the other three do not. The other two rows show the results of pooling all samples of the same tumor type together. Here the distinction usually fades, suggesting that usually the different types of rearrangement behave similarly within the same tumor type. Also, Notice that in different contexts all three factors plays a significant role (i.e. one factor is not always a byproduct of the other two).

Hypermutability near breakpoints

Analysis of the relationship between the sites of somatic mutations and rearrangements showed that the rate of somatic single nucleotide variations is significantly elevated near breakpoints (Figure 4-9, Methods Section). The effect can be detected in very close proximity to the breakpoint, but it becomes even stronger when calculated across 100bp – 1Kbp surroundings. Notice that the windows are non-overlapping, i.e. each window has a "hole" in the middle associated with the previous smaller window, and therefore the hypermutability is detectable also in regions far from the breakpoint. The increase in mutation frequency in a 1kbp window around breakpoints often reaches a staggering 100x – 3000x fold for several samples (Figure 4-10-A). The relationship between hypermutability and rearrangements was noted previously in various contexts[21], and we also previously showed it specifically for prostate cancer[7]. Here I demonstrate that this is true across many cancer types.

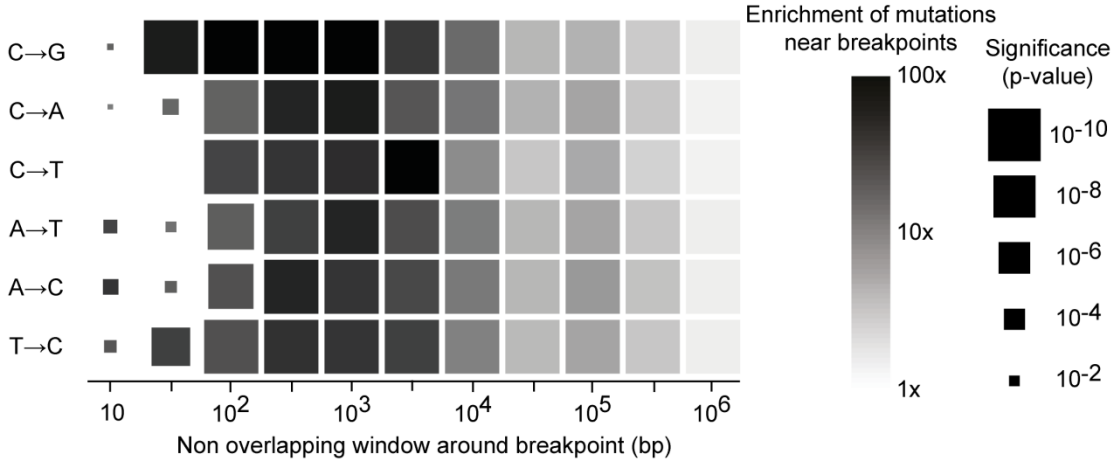


Figure 4-9 - **Enrichment of mutations across all samples by mutation type.** Squares represent mutation rate in concentric non overlapping exponential windows around each breakpoint, compared to overall mutation rate in the 71 samples cohort, aggregating them together. Size represents p-value, and color represents ratio. Only significant (FDR<10%) results are shown. Hypermutation can be seen in a close proximity of the breakpoint, but it is even stronger in 100bp – 1Kbp surroundings.

The hypermutation cannot simply be explained by rearrangement and mutations occurring in the same regions of the genome that are hyper-susceptible to all forms of genomic aberrations in all cases. I examined regions defined by the rearrangements of any given sample, and looked for mutations in those regions in all other samples of the same cancer type. While sometimes indeed elevated mutation rates can be noted (consistent with the hypothesis of fragile and hypermutable genomic regions), there

were almost always significantly more mutations (~16x increase in density) in samples identified by comparing to the genome-wide average (Figure 4-10-B).

The spectrum of the mutations surrounding breakpoints is significantly different than the spectrum over the entire genome, as can be seen in Figure 4-10-C, with $C \leftrightarrow G$ transversions being most highly enriched. $C \leftrightarrow G$ transversions were suggested to be caused by oxidative DNA damage[22, 23], and by base excision repair via uracil-DNA glycosylase and REV1 translesion synthesis[24, 25].

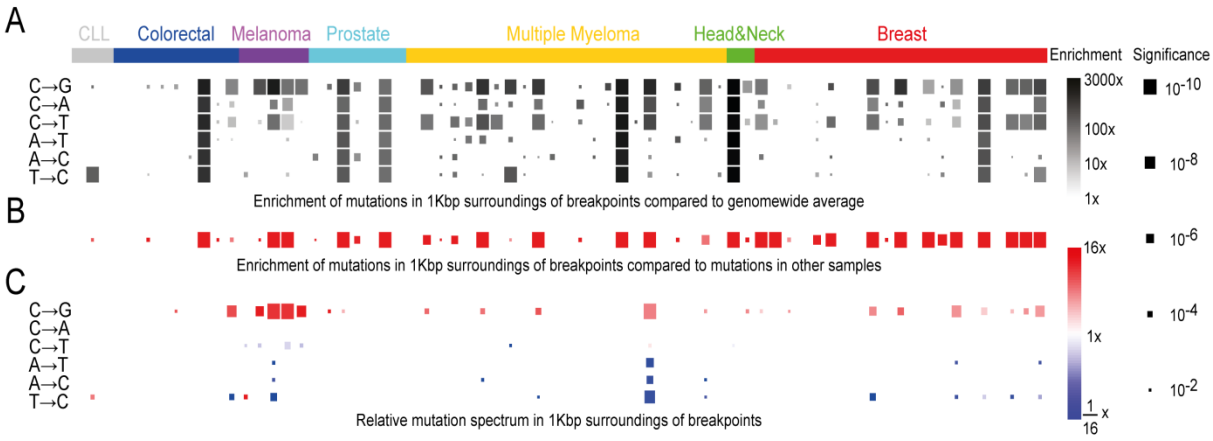


Figure 4-10 - **A. Enrichment of mutations across every sample in 1Kbp windows around breakpoints.** Size represents p-value, and color represent enrichment ratio. Only significant (FDR<10%) results are shown. For some samples the mutation rate can reach 1000x-3000x fold. **B. Hypermutation is not only due to rearrangement and mutations occurring in the same "bad" regions of the genome.** For each sample I defined the 1Kbp regions according to its rearrangement, and measured the mutations in those regions in all other samples of the same cancer type, aggregating them together. Squares represent p-value and ratio comparing the mutation rate in the selected samples to the mutation rate at the other samples of the same cancer type. Any sample with significant hypermutation displays significant elevation in mutation rate near breakpoints of that sample. **C. Mutation spectrum near breakpoints compared to spectrum across the genome of that sample.** Hyper mutated samples are often skewed towards C>G transversions near breakpoints. Melanoma samples show depletion of C>T transitions near breakpoints due to high C>T transitions across the genome.

$C \leftrightarrow G$ transversions are known to be enriched in breast cancer[26], where they tend to occur in a TpC (or GpA for $G \rightarrow C$) dinucleotide context. A similar context specific pattern also holds in lung cancer, ovarian cancer and melanoma[27, 28]. Mutations in that context are consistent with a DNA deamination by apolipoprotein B mRNA editing enzymes (*APOBEC1* and some of the APOBEC3 family genes)[29, 30]. I confirmed the enrichment of $C \leftrightarrow G$ transversions in the TpC context, but also observe that this

effect is significantly higher near breakpoints. Out of 25 samples that have more than 5 $C \leftrightarrow G$ transversions near breakpoints (1Kbp or less), 9 (5 breast cancer, 2 melanoma and 2 multiple myeloma) displayed significant enrichment of TpC context compared to transversions far from breakpoints (FDR<5%, Fisher's exact test).

One of the features of the translesion synthesis I suggest above is that it acts upon one of the strands and therefore only this strand will be mutated by the deamination. Indeed two multiple myeloma samples (MMRC0344 and MMRC0392) and four breast samples (BR-V-004, BR-V-006, BR-V-008 and BR-V-010) had a least one breakpoint with significant strand specificity (FDR<10%, see Methods).

Recently, clusters of mutations were discovered in breast cancer, a phenomenon termed kataegis[31]. Some of those clusters were colocalized with rearrangements. Nik-Zainal et al identified 5 mutational signatures by statistical inference, two of which (B and E) were found to be enriched in the kategis events. Signature E is mainly $C \leftrightarrow G$ transversions in TpC context, and signature B is a combination of $C \leftrightarrow G$ and $C \rightarrow T$ in TpC context. My results are consistent with their findings, namely hypermutation near breakpoints, enrichment of $C \leftrightarrow G$ mutations in TpC context and strand specificity.

Methods

BreakPointer Algorithm

A general sketch of the algorithm is shown in Figure 4-11 below and explained here in detail:

Approximate somatic rearrangement are predicted from tumor and matching normal. I have used dRanger (Lawrence MS et al. In preparation), but other paired end mapping predictors can be used[32-35].

BreakPointer fishes for reads that span both sides of a putative breakpoint (henceforth "split reads"). This is done by collecting all read-pairs with one read mapped to the proximity of the approximate breakpoint, while the other end is not mapped, mapped with mismatches at its tip, or clipped. The window around the approximate breakpoint is selected so that it will account for the fragment size of the library, the direction of the read, the read length, and for Δ_a, Δ_b the estimated error in the prediction. Δ_b is the error before the breakpoint, that is the side the supporting reads are on, and Δ_a is the error after the breakpoint, so that prediction x is transformed into predicted interval

[x-Δ_b, x+Δ_a]. *$\Delta_a \neq \Delta_b$* since it is likely that the breakpoint took place after the last supporting read, but less likely that it will be before the middle of that read (or before the end of the read if it is clipped). *$\Delta_a(n)$* and *Δ_b*(n) are functions of *n*, the number of supporting discordant pairs for the prediction, reflecting the confidence in the prediction. The parameters used in this analysis were *$\Delta_a(n \leq 6)=200bp$, $\Delta_b(n \leq 6)=80bp$, $\Delta_a(n>6)=60bp$, $\Delta_b(n>6)=40bp$*.

After the list of candidate split reads was compiled, the reads are aligned by a modified Smith-Waterman algorithm (see next section), that allows aligning each split read to the two regions in the reference genome, between which the rearrangement had occurred.

Each breakpoint is called by a majority vote among those reads that were aligned with good enough score to both sides. If 20 split reads which agree on the same breakpoint are found, further aligning is skipped to save running time. Split reads that are split in the first or last *$\beta=7bp$* are not counted, as they might be due to short similarities in the sequences, and don't constitute hard enough evidence. The confidence in the exact breakpoint is declared according to two parameters – the number of supporting split reads which agree on the same exact breakpoint, and their average alignment score.

Once the breakpoint is called, all the candidate split reads in the list are being aligned to the newly fused sequence using BWA[36]. This ensures finding the real number of supporting reads, as the alignment process was halted in the middle, and to salvage reads that align to the consensus breakpoint sequence, but were originally misaligned to suggest a wrong breakpoint. BreakPointer outputs the fused sequences together with all the split reads aligned to it, in the standard BAM format[37] containing the fused sequences with the split reads aligned to it, allowing convenient browsing using IGV[38] (see Figure 4-11-B) or similar tools. Moreover, a detailed report is produced containing all the breakpoints and for each breakpoint it reports the number of supporting reads and their average support quality, the rearranged sequence (before and after the breakpoint), as well as sequence that got dropped in the rearrangement or foreign sequence that was inserted (in case one exists) and the amount of homology between the two fused regions (both exact homology, and an alignment based score, allowing shifts, gaps and mismatches).

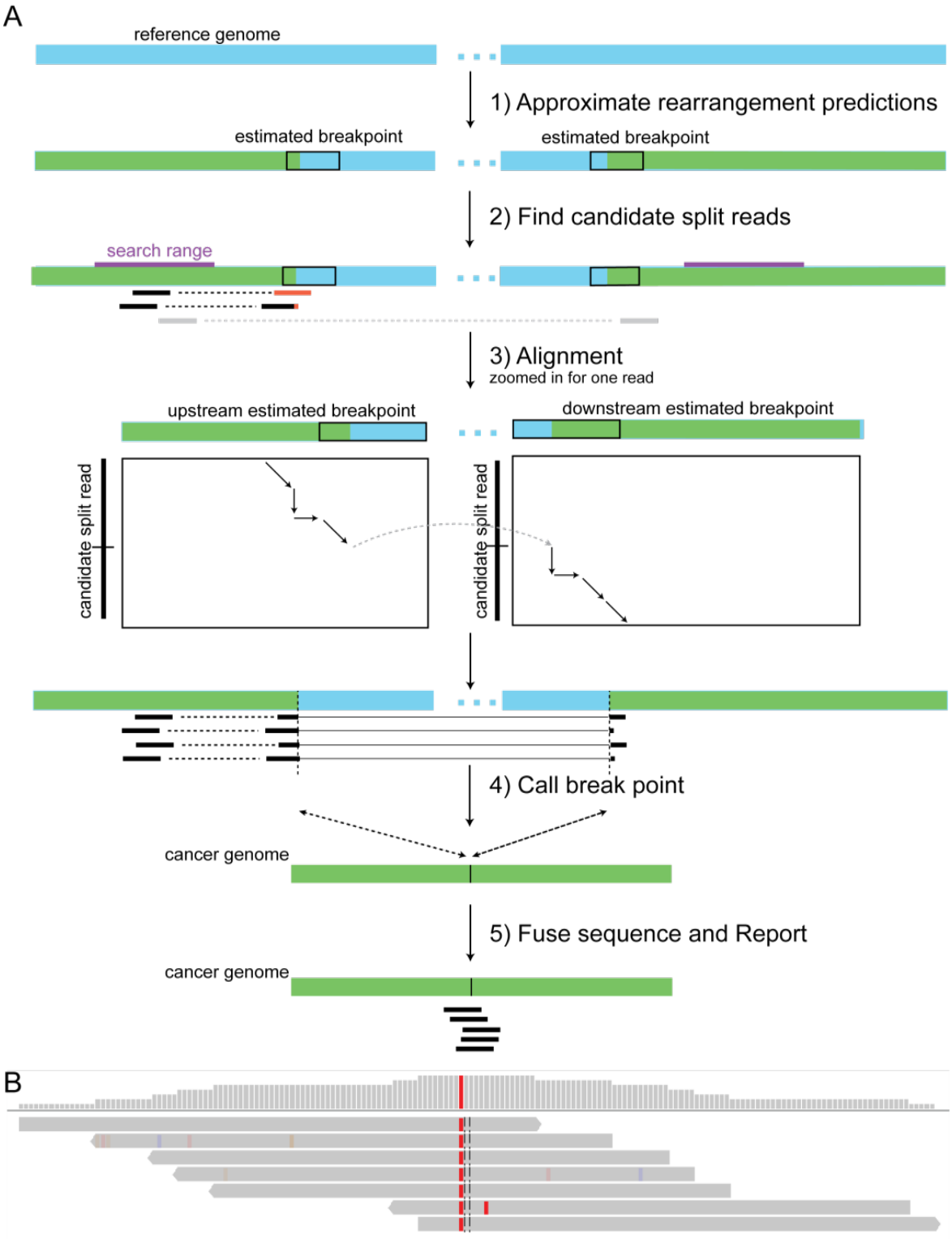


Figure 4-11 - **A. The five steps of the BreakPointer algorithm.** Green represents the cancer genome, as sampled from a specific tumor, blue the human reference genome. Due to a rearrangement, the two green parts are adjacent in the cancer genome, but not on the reference genome, and hence the blue bar is revealed. Reads in black are aligned reads, and in red unaligned or mismatches. **B. Actual screen shot of IGV** showing one rearrangement and its supporting split reads aligned to the rearranged cancer genome produced by BreakPointer. The breakpoint on the reads is in the middle of the screen shot marked by dashed black line. In this particular case there is a single mutation near the breakpoint, marked in red (representing the base T).

Modified Smith-Waterman split read alignment

The algorithm receives two regions of the genome between which the rearrangement is predicted to occur, and a read suspected to span the rearrangement. The output of the algorithm is where to best split the read into two parts so that the first part will map to the first region and the second to the second region, maximizing the alignment score. If the read is mapped just as well (or almost as well) without splitting it, the algorithm will declare that it is not a split read.

The algorithm is based on Smith-Waterman[39] approach, constructing two alignment matrices in a dynamic programming fashion, one for each region of the genome. The first matrix is rather straight forward:

$$M_1(i,j) = max\begin{Bmatrix} 0 & \\ M_1(i-1,j-1) + w(g_i, r_j) & \text{Match / Mismatch} \\ M_1(i-1,j) + w_{del} & \text{Deletion} \\ M_1(i,j-1) + w_{ins} & \text{Insertion} \end{Bmatrix}$$

Where $w(g_i, r_j) = \pm p$ is the base-calling error probability of the read at base j, $w_{del} = 3$ and $w_{ins} = 3$. After M_1 is constructed, if it attains a very high or low maximal value, the algorithm declares the read should not be split and finishes. Otherwise, it continues to construct a second matrix M_2 for the second sequence:

$$M_2(i,j) = \max\begin{Bmatrix} 0 & \\ M_2(i-1,j-1) + w(g_i, r_j) & \text{Match / Mismatch} \\ M_2(i-1,j) + w_{del} & \text{Deletion} \\ M_2(i,j-1) + w_{ins} & \text{Insertion} \\ \max\limits_{k} \; M_1(k,j-1) + w(g_i, r_j) + w_{split} & \text{Split} \\ \max\limits_{\substack{k \\ l<j}} \; M_1(k,l) + w(g_i, r_j) + w_{fs} & \text{Foriegn Sequence} \end{Bmatrix}$$

Where $w_{split} = -4$ and $w_{fs} = -7$. Hence, M_2 allows jumping from a previous point in M_1, with some penalty, and possibly open a gap at constant penalty, to allow an insertion of foreign sequence. If one of these options was chosen, the jump coordinates (k and l) are kept to allow trace back. If M_1 attains a larger maximum than M_2 BreakPointer deduce there's no breakpoint. Otherwise, M_2 is traced back to locate the breakpoint (both on the read and in the genome). The quality of the

alignment is taken as the maximal value of M_2, divided by the read length (for standardization).

Parameter Optimization and Manual Review

Free parameters of BreakPointer were optimized using data from two samples- PR-2832 and PR-1783, along with their 454 validation data. Each set of parameters was optimized separately. First, the parameters of the read fishing (that set the size of the window around the approximate prediction, as specified above) were set to capture most of the split reads, without adding too much "noise" (reads that are not split over the breakpoint). Afterwards, the parameters of the split read alignment ($w_{del}, w_{ins}, w_{split}, w_{fs}$) were selected based on their internal logic and a few selected study cases of rearrangements and split reads of different characteristics, so that the split reads will align correctly. Finally, parameters of the rearrangement calling (number of reads required, minimal alignment quality and β) were selected to optimize specificity and sensitivity of the two samples.

300 declared breakpoints have been manually reviewed with IGV. To allow quick review a tool was written to control three IGV instances, so that for each rearrangement, one IGV instance will display the split reads aligned to the fused sequence (the BAM file BreakPointer produces, see Figure 4-11-B), and two more IGV instances will focus on either breakpoint loci to show the reads that were not split mapped to the genome in those regions (the original paired end BAM file). Out of the 300, only one was clearly wrong according to the available data (either there is no breakpoint at all at these coordinates, or the wrong breakpoint was called).

dRanger Algorithm

Chromosomal rearrangements are identified by a method developed by the Broad's Cancer Genome Analysis group, called dRanger (Lawrence M.S. et al., manuscript in preparation, http://www.broadinstitute.org/cancer/cga/dRanger). First, discordant read pairs are identified in the tumor. These are read pairs that map to different chromosomes or in unexpected positions (>600bp apart) or unexpected orientations (incorrect order on opposite strands or any order on the same strand) on the same chromosome. Second, clusters of discordant pairs are used to nominate potential rearrangement events. Candidate rearrangements are removed if there are any supporting discordant pairs for the same event in its corresponding matched normal or

in a panel of additional normal genomes sequenced at the Broad Institute. Third, a series of additional filtering metrics is computed for each candidate rearrangement: (1) the fraction of nearby reads with a mapping quality of zero; (2) the number and diversity of other discordant pairs in the vicinity of the breakpoints; and (3) the standard deviation of the starting positions of the supporting read pairs. These filtering metrics are combined into an overall quality measure (0 to 1), which serves as a multiplicative scaling factor to convert the number of supporting read pairs to a score for the rearrangement. Based on independent validation experiments, I consider rearrangements with a score of 4.0 or higher as real. As validation data is partial and it is hard to tell apart false positives from problems in the validation step (such as complex rearrangements or PCR failure), I used all predictions for the analysis presented here (regardless if they were validated or not). To test whether any of the main conclusions is an artifact generated by false positives I repeated the analysis only with the validated results. This analysis did not alter the results significantly (data not shown). Approximate locations of rearrangement breakpoints are assigned based on the boundary of all reads in supporting read pairs.

Implementation and Availability

The different parts of BreakPointer are implemented in Matlab (general framework), C (Modified Smith-Waterman alignment for processing efficiency) and Java (to efficiently interact with SAM-JDK I/O operations). BreakPointer is incorporated into the Broad Institute Cancer Genome Analysis pipeline ('Firehose'), and is therefore available as a GenePattern module[40].

The input to BreakPointer consist of a BAM file with the paired end reads, blacklist of possibly erroneous lanes, the version of human genome assembly, the size of each job for a parallel run, and a text file that describe the PEM-based breakpoint predictions. The text file can be a standard dRanger output file, or a tab delimited text file with a header row, containing 'chr1','chr2','pos1','pos2','str1','str2' columns which represent the predicated rearrangement (chromosome, position, and strand for the two breakpoints) and a 'tumreads' column which represent how many reads support the prediction.

BreakPointer is available at http://www.broadinstitute.org/cancer/cga/BreakPointer.

BreakPointer Sensitivity analysis

In order to determine what is the required physical coverage that enables to pinpoint breakpoints, I down-sampled data with 10 or more supporting reads, and measured the rate of affected breakpoints. With 3 or more supporting reads, 98% of the breakpoint calls were the same; with 5 reads or more, all breakpoint calls were identical (Figure 4-12).

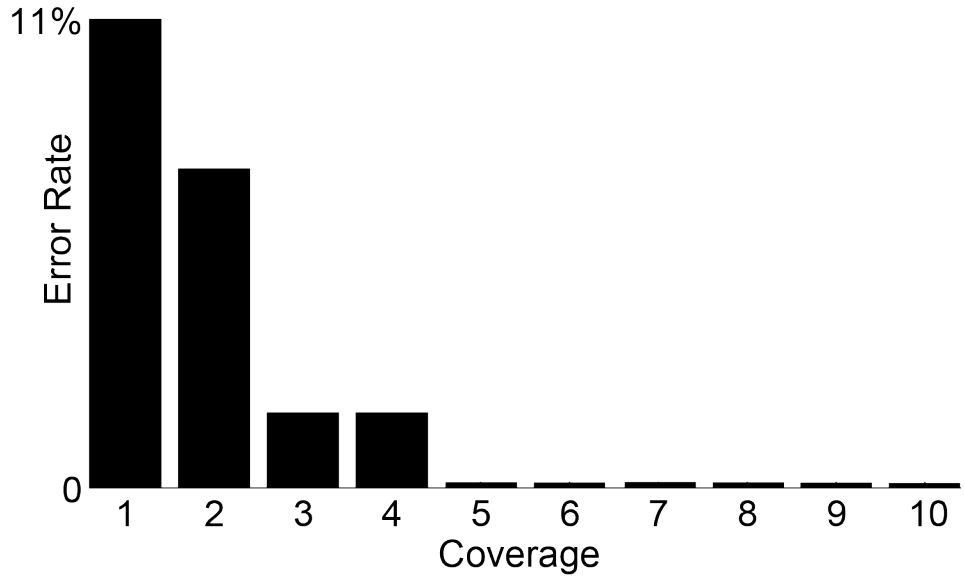


Figure 4-12 – **Down-sampling to quantify the required coverage**. The fraction of breakpoints predicted incorrectly out of all CRC breakpoints with at least 10 supporting reads, as a function of the number of reads randomly selected to recall the breakpoint (compared to the call made with all supporting reads).

Data and preprocessing

The data used for the analysis was whole genome shotgun sequencing preformed as described in references[7, 8]. Candidate chromosomal rearrangements were identified from the observation of multiple discordant read pairs using dRanger (see below). BreakPointer was originally designed to use MAQ[41] alignments, but was also adapted to BWA[36]. BWA introduced later on advanced clipping features making the identification of split reads easier, allowing use of alternative rearrangement detection algorithms such as CREST[42]. Breast and head-and-neck samples were aligned using BWA, all the other samples using MAQ.

Validation

Rearrangements predicted by dRanger (with at least 3 supporting discordant reads) were validated by PCR followed by pooled 454 sequencing. PCR primers were

designed using Primer3[43] such that they spanned the predicted chimeric junction and would produce an amplicon approximately 300–350bp long. PCRs were performed on whole genome amplified product for both tumor and normal DNA (For somatic breakpoints, only the tumor DNA would be expected to yield a product). Each PCR product was quantified using a NanoDrop Spectrophotometer (Thermo Scientific, Wilmington, DE). PCR products were pooled such that: (1) equal amounts of tumor products were combined, (2) the same volumes were taken from the corresponding normal products, and (3) matching tumor and normal products were placed in separate pools. Libraries for 454 sequencing were prepared from each pool and sequenced in separate regions of a 454 Genome Sequencer FLX System (454 Life Sciences, Branford, CT). Primer sequences served as unique barcodes for identifying the source PCR product for each 454 read. A rearrangement was judged to be somatic if the predicted chimeric product was detectable in tumor DNA and not normal DNA.

To validate BreakPointer results, the fused sequence generated by BreakPointer was aligned by Smith-Waterman[39] to all the sequences of the appropriate amplicons (or their reverse compliment). For each amplicon, the alignment was declared to be successful, if it contained no gaps in a 20bp window around the breakpoint (to ensure exact pinpointing) and at least 95 matches in a 100bp window. Notice that since BreakPointer fuses the reference genome, some mismatches with cancer genomes are expected (due to germline and somatic point variations).

Statistical analysis of microhomologies

Wilcoxon rank-sum test was used to compare the observed microhomology distribution to the expected background, for each type of rearrangement separately. The background for each test is based on hypothetical rearrangements constructed by taking all possible breakpoint pairs among the breakpoints belonging to a particular rearrangement type, and then computing the distribution of microhomologies in this set of hypothetical rearrangements. To evaluate the difference between the histograms of the different rearrangement types, I used Scholz-Stephens' k-sample Anderson–Darling statistic[44] to measure the similarity between the histograms. I then tested the significance of this value based on 10^6 sets of 'permuted' histograms generated under the null hypothesis in which the histograms are in fact not different. To generate each set of histograms, I randomly permuted the observed microhomology among the five rearrangement types. I then computed the Anderson-Darling statistic for each set, and

the p-value is simply the fraction of sets with higher or equal Anderson-Darling statistic than the original five histograms. To evaluate the contribution of the short deletions and the tandem duplications to the significance, I repeated the analysis omitting one or both. Excluding the short deletions and keeping the tandem duplications yields histograms that are still significantly different ($p<10^{-6}$). However, when removing the tandem duplications and keeping the short deletions the results are less significant p=0.03 and when omitting both the histograms are no longer significantly different (p=0.15).

To evaluate significant deviations from expected distribution by composition of rearrangements, I compared the observed average microhomology to a background distribution controlled for the different types of rearrangements, and calculated empirical p-values. I constructed the background distributions of average microhomology for each sample, by sampling 10^6 times the appropriate number of rearrangements of each type. Microhomology was capped at 6bp, to eliminate the unwanted effect of few rearrangements with large homology. Similarly, for the cancer type specific analysis, all samples of the same cancer-type were pooled together and deviations from the appropriate background of the pool were calculated.

Usually microhomology is defined by perfect homology. However, biological mechanisms mediating microhomology might induce imperfect homology (i.e. sequence similarity of less than 100%). To assess the sensitivity I also attempted to define rearrangement with microhomology by requiring at least 5 matches and up to 2 mismatches (i.e. sequence similarity > 71%). Only 12% of those rearrangements did not have 2bp perfect microhomology, yielding no significant change in any aspect by adding this new definition. No correlation (ρ=-0.009, p=0.53) was detected between microhomology and coverage, excluding the possibility of a detection bias due to coverage.

Breakpoint distribution statistical analysis

Enrichment and depletion of breakpoints in different regions of the genome, defined by replication time, GC content and distance to transcribed gene, were computed by random generated distributions controlled for chromosome and coverage. First nearby breakpoints (up to 2500bp away) were consolidated into a single “event”. This was needed since nearby breakpoints were probably a result of one DNA breakage event. Controlling for chromosome was required to avoid artifacts resulting just from the

chromosome identity. These steps are specifically important for short deletions and in the presence of complex events (such as balanced translocations (Berger et al. 2011) or variants of chromothripsis[45]) that occur in several of the samples, as we previously reported[7, 9, 12]. Controlling for coverage (using the average coverage of all samples aligned to the same genome build) was needed because the ability to detect rearrangements depended on coverage (Figure 4-13).

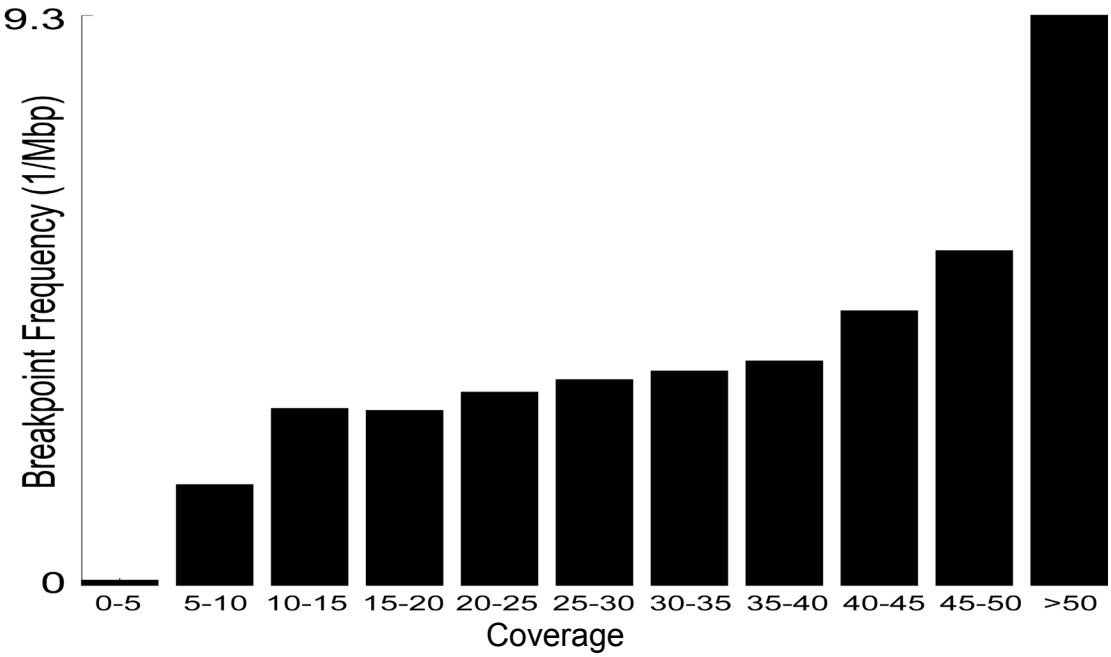


Figure 4-13 – Frequency of breakpoints summing over all 71 samples, as a function of the average coverage.

For each event, 100,000 locations (one per iteration) were generated uniformly from all locations on the same chromosome having the same coverage (in bins of 5, average coverage of all samples using the same genome build). The genome was binned to the following bins: low GC (0-36%), medium GC (36%-45%) and high GC (45%-100%). Transcribed regions (transcribed gene in 100Kbp), medium (100Kbp-500Kbp) and untranscribed regions (no transcribed gene in 500Kbp) (see below definition of transcribed gene). Replication time was binned according to late/early ratio[46] at $(-\infty,-0.8],(-0.8,0]$, $(0,0.8],(0.8,\infty)$. Changing the thresholds did not affect the essence of the results, other than losing sensitivity for too small or too big bins (data not shown). For every bin I counted the number of breakpoints for both the observed breakpoints and the random breakpoints. All of these counts were used to compute nonparametric p-values (observed rates). Enrichment or depletion was determined by

picking the lower of the one-sided p-values, and p-values were then corrected for multiple hypotheses by the Benjamini-Hochberg FDR procedure[47]. Logistic regression was used to study the contribution of each parameter to the joint distribution, assuming the distribution of the number of breakpoints that fall in each bin was binomial. The predictor variables are the %GC, transcription and replication time of each bin, and the predicted response variable is whether a genomic locus that falls in that bin breaks or not. The genome was sampled exactly at the loci where the permuted breakpoints from the enrichment test fell.

Transcribed genes were identified by picking the 10,000 most expressed genes on average from a matching dataset, as described below. DNA replication time data for H7 hESC cells was obtained from Ryba et al.[46], remapping to hg19 build was done via UCSC genome browser's tool liftOver. The GC content was called in 100Kbp windows. The only noticeable effect of using 10Kbp or 1Mbp windows was equivalent to slightly shifting the bin thresholds (as the smaller the windows size is, the more disperse the %GC distribution).

To look for genes mutated in LLU or EHT samples, I examined all mutations within genes, other than silent mutations or mutations in introns (but including mutations in promoters and UTR). I choose only genes which have the potential to be differentially mutated, that is that are mutated in at least 3 samples and are not mutated in at least 3 samples within LLU and EHT samples. Fisher's exact test was used to calculate the probability of a gene to be mutated in as many LLU or EHT samples.

Transcription data for the different tumor types

ExpO data (https://expo.intgen.org/expo/public/) is used to deduce active genes in prostate, breast and colorectal (by averaging colon and rectal samples) tissues. Multiple myeloma genes were deduced from reference [48] as downloaded from GEO website (GSE2658). Melanoma genes were deduced from reference [49] as downloaded from GEO website (GSE7127). Head and neck cancer genes were deduced from reference [50] as downloaded from GEO website (GSE6791). CLL genes were deduced from reference [51] as downloaded from GEO website (GSE13159). The same analysis have been repeated when transcribed regions of the genome for all cell types were deduced from ChIP-seq of H3K3me3 marks of human embryonic stem cells by Broad/Bernstein lab, as downloaded from the UCSC ENCODE webpage[52]. As expected the results of this analysis were very similar (results not shown).

Mutation rate statistical analysis

To test for enrichment of mutations near breakpoints, the same generated background distribution described above was used to count how many breakpoints had at least one mutation in any given window around the breakpoint. As breakpoints with a nearby mutation are rare events, Poisson distribution was assumed to infer p-values. When comparing to several samples together (Figures 4-10-A and 4-10-C above) mutations were aggregated into one virtual sample with all the mutations. To test for the enrichment / depletion of transitions and transversions near breakpoints, I performed a Fisher's exact test for each sample, on the number of mutations of each type near breakpoints, versus their distribution over all the genome. A similar Fisher's exact test was used to compare $\mathrm{TpC} \rightarrow \mathrm{TpG}$ out of all $\mathrm{C} \leftrightarrow \mathrm{G}$ transversions, near breakpoints and over all the genome. Fisher's exact test was also used to compare the mutation enrichment near breakpoints with the enrichment of mutations in other samples of the same tumor type in the same regions. To compute the frequency of mutations over all the genome and near rearrangements, mutations of each type were counted and divided by the total number of base pairs of the appropriate type that were covered enough to call mutations for.

To estimate the strand specificity of mutations near breakpoints, I examined all the 10Kbp windows around breakpoint that had at least 15 mutations. Mutation rate in the window was calculated on both strands (e.g. $\mathrm{C} \rightarrow \mathrm{T}$ and $\mathrm{G} \rightarrow \mathrm{A}$ together), and then binomial distribution was used to estimate the probably of having as many mutations on a single strand in that window (e.g. either $\mathrm{C} \rightarrow \mathrm{T}$ or $\mathrm{G} \rightarrow \mathrm{A}$).

Discussion

I identified three genomic factors that significantly affect, in a sample specific manner, the distribution of breakpoints: GC content, transcription, and replication time. The scales on which transcription affects the distribution of breakpoints suggest that the main effect is through the 3D DNA structure of the genome, i.e. the different open/closed chromatin compartments (present mostly during interphase). DNA replication time suggests co-localization, mostly during replication[46, 53], and was shown to affect rearrangements in bacteria[54, 55] and has been recently suggested for cancer as well[56]. GC content might affect breakpoint distribution by sequence-dependent mechanisms (such as homology), or may simply be correlated to other

biologically relevant factors. I show that the three factors, although highly correlated, are not redundant, and each may contribute differently in different contexts, e.g. different samples and different rearrangement types. We previously showed some correlation between breakpoints and transcription for prostate samples[7]; here I extend our analysis and offer a possible explanation. This genomic scale is consistent with recent discoveries that during interphase, transcription occurs in distinct compartments in the nucleus, and that untranscribed regions occupy other compartments[57-60]. It is known that often breakpoint-pairs of individual rearrangements occur in nearby segments of the DNA[56, 61-64]. However, I find that many breakpoints, belonging to different rearrangements, also tend to occur in some samples in transcribed/early replicating compartments and in other in untranscribed/late replicating compartments. This is not only an artifact of the breakpoint-pairs of individual rearrangements, as a similar pattern is observed when selecting randomly only one breakpoint of each rearrangement and repeating the analysis (Figure 4-14). This observation is consistent with a model in which one or more events had occurred, each causing several breakpoints within the same compartment, perhaps due to a strong DNA damaging event (as suggested to cause chromothripsis[45] when occurring during metaphase). Incorrect fusion of the resulting nearby fragments then yields the observed rearrangements. Moreover, I suggest *APC* deficiency as a mechanism that may contribute to DNA breakage in late replicating, low %GC, untranscribed regions of the genome, during or after mitosis. However, the accuracy of my findings regarding transcription and replication may be imperfect, as I did not measure transcription and replication in the specific tumor samples analyzed here. Reassuringly, time of replication is mostly constant in different tissues[65]. To make sure that the small difference in the patterns of transcription does not have a big impact, I have deduced the expression profile for each type of cancer separately. Moreover, repeating the analysis with different expression profiles yielded similar results (data not shown).

This data-driven model of the breakpoint distribution is not predictive at this point, and requires the full analysis of breakpoints in each sample. Due to the complexity of the effects, I believe such an approach is necessary to assess the significance of driver rearrangements across cancer. Since the cohort size is still a limiting factor, statistical inference of the causes of the different behavior of different samples is not yet possible. However with the large number of cancer whole genome sequences

becoming available, this is expected to change in the near future, allowing similar methodology to provide an understanding of different biological processes that contribute to the variability across samples and types of alterations.

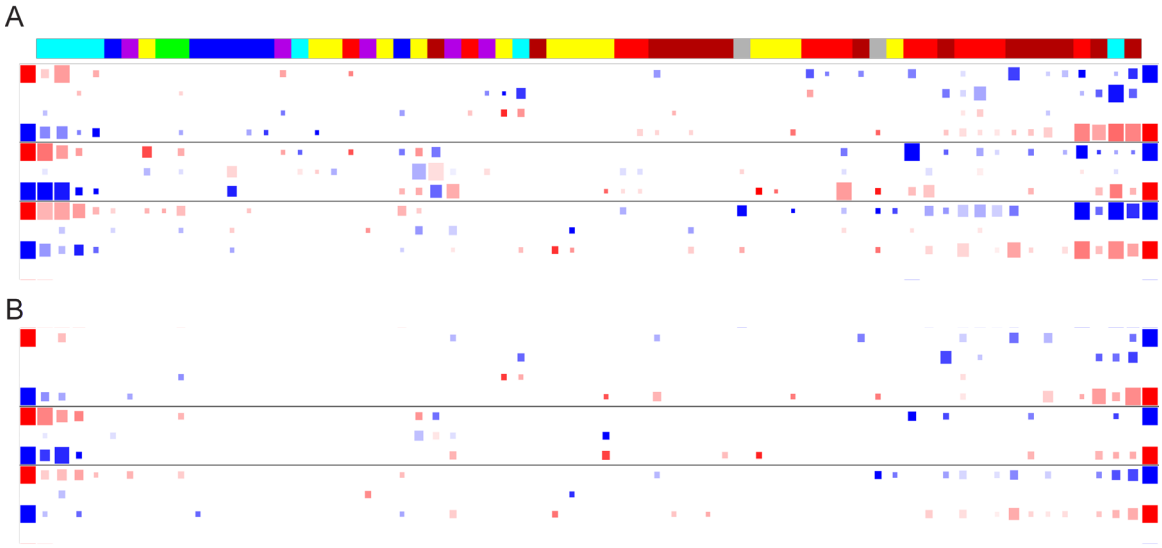


Figure 4-14 – Breakpoint distribution as function of transcription, replication and GC content across samples, as in Figure 4-6-A, keeping the same ordering, showing all results > 20% FDR, but when: Choosing only one breakpoint of any rearrangement(**A**) or consolidating every two breakpoint at most 1Mb apart (**B**). Notice that the patterns are similar.

Integrative analysis of mutations and exact breakpoint loci revealed a global hypermutability near breakpoints, common to almost all samples. I suggest that the hypermutability might be related to base excision repair caused by *APOBEC* deamination that can cause both DNA breakage and the mutations I observe near breakpoints. Moreover, the strand specific pattern seen in some of the samples may suggest that it is caused by translesion synthesis, that is known to occur in base excision repair. This emphasizes the complexity of understanding the deterioration of genome stability, the effect of different DNA repair mechanisms, and the need to integrate the different data types in order to better understand them. It also has a practical impact on modeling of background mutation rates in cancer. The different mutation spectrum near rearrangements suggests that different mechanisms generate or repair these mutations, and may help identify these mechanisms. Further study is required to understand the relationship between breakpoints and the processes that govern mutation spectra.

5. Analysis of rearrangements across different cancers

In collaboration with members of the Cancer Program at the Broad Institute of MIT and Harvard. Publications #4,#5,#7 amd #9.

Introduction

In this chapter I describe the application of dRanger and BreakPointer, described in chapter 4, to the analysis we performed on several different cancer datasets. This work was done in collaboration with members of the Cancer Program at the Broad Institute, and especially Mike Berger, Mike Lawrence, Adam Bass, Gad Getz, Levi Garraway and Matthew Meyerson. As I was involved only in the analysis of rearrangements, I will describe here only that, more details are in our publications[7-9, 12] (appearing as publications 4,5,7 and 9 of the list presented in Chapter 9).

Methods

The complete genomes of tumors and patient-matched normal samples were sequenced to approximately 30-fold haploid coverage on an Illumina GA II sequencer. DNA was extracted from patient blood and from tumors following radical prostatectomy, and was subjected to extensive quality control procedures to monitor DNA structural integrity, genotype concordance, and tumor purity and ploidy. Standard paired-end libraries (~400-bp inserts) were sequenced as 101-bp paired-end reads. Raw sequencing data were processed by Illumina software and passed to the Picard pipeline, which produced a single BAM file for each sample, storing all reads with well-calibrated quality scores, together with their alignments to the reference genome. BAM files for each tumor/normal sample pair were analyzed by the Firehose pipeline to characterize the full spectrum of somatic mutations in each tumor, including base pair substitutions, short insertions and deletions, and large-scale structural rearrangements. A subset of base pair mutations and rearrangements were validated using independent technologies in order to assess the specificity of the detection algorithms. FISH was also performed for selected recurrent rearrangements. **A complete description of the materials and methods is provided in our papers mentioned above.**

Co-occurrence of selected transcription factor binding sites and chromatin marks with breakpoint locations and mutations in prostate cancer

For each prostate cancer genome (as well as published melanoma, lung, and breast cancer genomes), we tested whether the associated rearrangement breakpoints occurred closer to or farther from a given set of ChIP binding sites than expected by chance. We downloaded pre-computed ChIP-Seq binding peaks for the following transcription factors and chromatin marks in the androgen-sensitive, TMPRSS2-ERG fusion positive prostate cancer cell line VCaP: AR (followed by treatment of R1881, a synthetic agonist of the androgen receptor), ERG, RNA polymerase II, acetylated histone H3, trimethylated histone H3K4, trimethylated histone H3K36, trimethylated histone H3K9, and trimethylated histone H3K27[66]. The number of peaks in each experiment ranged from 1,725 to 42,568. We also downloaded pre-computed genome-wide ChIP-Seq binding peaks for AR, H3K4me3, H3K36me3, H3K9me3, and acetylated histone H3 in the ETV1+ prostate cancer cell line LNCaP[66]; for AR in the ETS- prostate cancer cell line PC3[67]; for H3K4me3, H3K36me3, and H3K27me3 in 3 cell lines from the ENCODE project[52] (GM12878, K-562, and H1ES); and ChIP-chip binding peaks for estrogen receptor (ER) in the breast cancer cell line MCF7[68]. In addition to the prostate cancer rearrangements presented here, we considered pre-computed rearrangements for published genomes in a melanoma cell line[69], a small cell lung cancer cell line[70], a primary lung cancer[71], and 24 breast cancer cell lines and primary tumors[4]. (We later discarded 6/24 breast cancers with fewer than 20 rearrangements.)

To test for enrichment or depletion of a prostate tumor's rearrangements near a given set of ChIP-Seq peaks, we calculated the rate of breakpoints within the aggregate of all sequence intervals ±50Kbp surrounding each peak. This was compared to the background rate of breakpoints, which we estimated by taking the average of 1,000 simulations in which we controlled for coverage and structure. Simulated breakpoints were randomly generated at positions matched in sequence coverage to the observed breakpoints, to control for hidden correlations between breakpoints and ChIP-Seq peaks due to sequencing bias. (Background sampling considered the mean sequence coverage across all 7 prostate genomes in bins of size 5 (top bin ≥ 50-fold depth).) To control for structure, simulated breakpoint pairs corresponding to intrachromosomal rearrangements were preserved at fixed distances such that one end was perfectly matched in sequence coverage and the other end occurred at a site with no less than

(but possibly greater than) the observed sequence coverage. (Controlling for structure was necessary to account for non-independent events from small intrachromosomal inversions and deletions.) Significance of enrichment or depletion of observed breakpoints compared to background was calculated according to the binomial distribution. In addition to a binomial p-value, we also computed the ratio of the observed rate to the background rate to determine the effect size independent of the total number of rearrangements detected in a given sample.

We repeated this calculation using different window sizes and found that the effects were consistent for intervals ranging from ±1 Kbp to ±1 Mbp surrounding each ChIP-Seq peak. To be sure that significant associations were not the result of a small number of driver genes, we repeated calculations upon removing all rearrangements involving any of 16 genes we found to be recurrently disrupted in the 7 prostate tumors presented here (57 rearrangements total).

To test for enrichment or depletion of point mutations near a given set of ChIP-Seq peaks, we repeated the calculations exactly as described above using a coverage-matched simulated background. (We did not attempt to preserve the distance between mutations on the same chromosome.) To test for enrichment or depletion of point mutations near rearrangements in the corresponding prostate genome, we repeated the calculations using a coverage-matched simulated background in different window sizes surrounding each breakpoint.

PCR and massively parallel sequencing of rearrangements in prostate cancer

Predicted rearrangements were validated by PCR followed by pooled 454 sequencing. PCR primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3) such that they spanned the predicted chimeric junction and would produce an amplicon approximately 300–350bp long. PCRs were performed on whole genome amplified product for both tumor and normal DNA. (For somatic breakpoints, only the tumor DNA would be expected to yield a product.) Each PCR product was quantified using a NanoDrop Spectrophotometer (Thermo Scientific, Wilmington, DE). PCR products were pooled such that: (1) equal amounts of tumor products were combined, (2) the same volumes were taken from the corresponding normal products, and (3) matching tumor and normal products were placed in separate pools. Libraries for 454 sequencing were prepared from each pool and sequenced separate regions of a 454 Genome Sequencer FLX System (454 Life Sciences, Branford, CT). Primer

sequences served as unique barcodes for identifying the source PCR product for each 454 read. A rearrangement was judged to be somatic if the predicted chimeric product was detectable in tumor DNA and not normal DNA.

Fluorescence in situ hybridization (FISH) for MAGI2, PTEN, CADM2, and CSMD3 rearrangements in prostate cancer

To assess the status of PTEN, we used a locus specific probe and a reference probe. To assess for inversion of the MAGI2 gene, a unique FISH assay was designed. Probes spanning the ends of the gene were labeled red (3' end) or green (5' end). A third probe, also labeled green, acted as a reference for the arrangement of the gene. A chromosome with no gene inversion showed a red signal (3' end of MAGI2) followed by two green signals (5' end of MAGI2 then the reference probe at 7q36). A chromosome with the gene inversion showed the red signal between the two green ones, indicating that the 3' end and the 5' end have been inverted. To identify rearrangements disrupting CADM2 and CSMD3, we utilized break apart FISH assays with probes positioned on both sides of the gene.

To determine the prevalence of each class of rearrangement, we surveyed an independent cohort of 90 patients (mean age 63.2 years) who underwent radical prostatectomy at Weill Cornell Medical College (New York, NY) as a monotherapy. The pathological stages ranged from organ confined to cases with extra-prostatic tumor extension.

Validation of the VTI1A-TCF7L2 fusion in colorectal cancer

To test the functional importance of the VTI1A-TCF7L2 fusion, we sought a cell line harboring this event. Because the fusion in CRC-9 is caused by a ~540Kbp deletion between VTI1A and TCF7L2, we studied SNP array data from 38 colorectal cancer cell lines to search for a similar deletion. We found that the cell line NCI-H508 carries such a deletion, and we showed the presence of an in-frame fusion transcript linking exon 2 of VTI1A to exon 5 of TCF7L2. We designed RNA-interference vectors targeting the sequence spanning the fusion. Two vectors that reduced the expression of the fusion mRNA by >70% as gauged by quantitative RT-PCR caused a dramatic reduction in the anchorage-independent growth of cells from NCI-H508 but not DLD-1, a colorectal cancer cell line that does not harbor the fusion gene. This result shows that the VTI1A-TCF7L2 fusion plays a critical role in NCI-H508 cell growth.

The NCI-H508 cell line was identified from SNP-array–derived copy number from a collection of 38 colorectal cancer cell lines in the Broad-Novartis Cell Line Encyclopedia. RNA prepared from samples of fresh-frozen colorectal adenocarcinomas or, in the case of the NCI-H508 cell line, a fresh cell pellet, were used for cDNA synthesis with the QIAGEN QuantiTect kit. cDNA quality was assessed by the ability to PCR amplify the GAPDH transcript. Passing cDNA was evaluated with a first round of PCR using primers to the 5′ untranslated region of VTI1A and exon 6 of TCF7L2 and then nested PCR using primers from the first exon of VTI1A and exon 5 of TCF7L2. Bands were gel purified, cloned (TOPO TA Cloning; Invitrogen) and sequenced to validate the presence and frame of fusion.

Results

Prostate cancer rearrangements analysis

The first and most intensive analysis was of prostate cancer. In prostate cancer rearrangements play a major role, and therefore we have performed a detailed analysis. Seven prostate samples have been whole genome sequenced and analyzed for mutations and rearrangement.

Careful analysis of the rearrangements revealed three main discoveries:

1) Complex patterns of balanced rearrangements - a distinctive pattern of balanced breaking and rejoining not previously observed in solid tumors: several genomes contained complex inter- and intra-chromosomal events involving an exchange of 'breakpoint arms'. A mix of chimaeric chromosomes was thereby generated, without concomitant loss of genetic material, that is, all breakpoints produced balanced translocations. See Figure 5-1.
2) Association of rearrangements and epigenetic marks - The location of rearrangement breakpoints from the TMPRSS2–ERG fusion-positive tumor PR-2832 showed significant spatial correlation with various marks of open chromatin in VCaP cells. These marks included ChIP-seq peaks corresponding to RNA polymerase II (pol II, $p=1.0 \times 10^{-15}$), histone H3K4 trimethylation (H3K4me3, $p=3.1 \times 10^{-7}$), histone H3K36 trimethylation (H3K36me3, $p = 3.5 \times 10^{-12}$) and histone H3 acetylation (H3ace, $p=9.5 \times 10^{-12}$). Similar statistical correlations were observed for peaks corresponding to the androgen receptor (AR) ($p=1.1 \times 10^{-5}$) and ERG binding sites ($p=4.9 \times 10^{-14}$), consistent with the

substantial overlap between AR and ERG binding locations in VCaP cells[66]. Rearrangement breakpoints from all four ETS fusion-negative tumors were inversely correlated with these same marks of open chromatin and AR/ERG binding. In fact, breakpoints from two of four ETS-negative tumors were significantly correlated with marks of histone H3K27 trimethylation (H3K27me3) in VCaP cells, which denote inactive chromatin and transcriptional repression. This result suggested that somatic rearrangements might occur within closed chromatin in some tumor cells, or that the epigenetic architecture or transcriptional program of some TMPRSS2–ERG fusion-positive cells differs markedly from that of ERG fusion-negative cells. In support of the former, we observed a similar enrichment of PR-2832 rearrangements and depletion of fusion-negative rearrangements near marks of active transcription profiled in several additional cell lines, including fusion-negative prostate cancer cell lines LNCaP and PC-3 as well as three cell lines derived from non-prostate lineages. Significance is calculated by random permutations of the breakpoints, controlled for coverage and chromosome identity. See Figure 5-2.

3) Recurrent rearrangements involving CADM2, PTEN and MAGI2 - In addition to the well-known TMPRSS2-ERG fusion, in 3 of the 7 tumors we detected rearrangements in CSMD3 and CADM2. These genes were rearranged at a frequency beyond that expected by chance, even after correcting for gene size. CSMD3 encodes a giant gene that contains multiple CUB and sushi repeats. CADM2 encodes a nectin-like member of the immunoglobulin-like cell adhesion molecules. Several nectin-like proteins exhibit tumor suppressor properties in various contexts. Two prostate tumors contained breakpoints within the PTEN tumor suppressor gene. In both cases, the rearrangements generated heterozygous deletions that were confirmed by FISH analysis. Two additional tumors harbored rearrangements disrupting MAGI2, which encodes a PTEN-interacting protein. Whereas both PTEN rearrangements involved chromosomal copy loss, the MAGI2 rearrangements were balanced events.

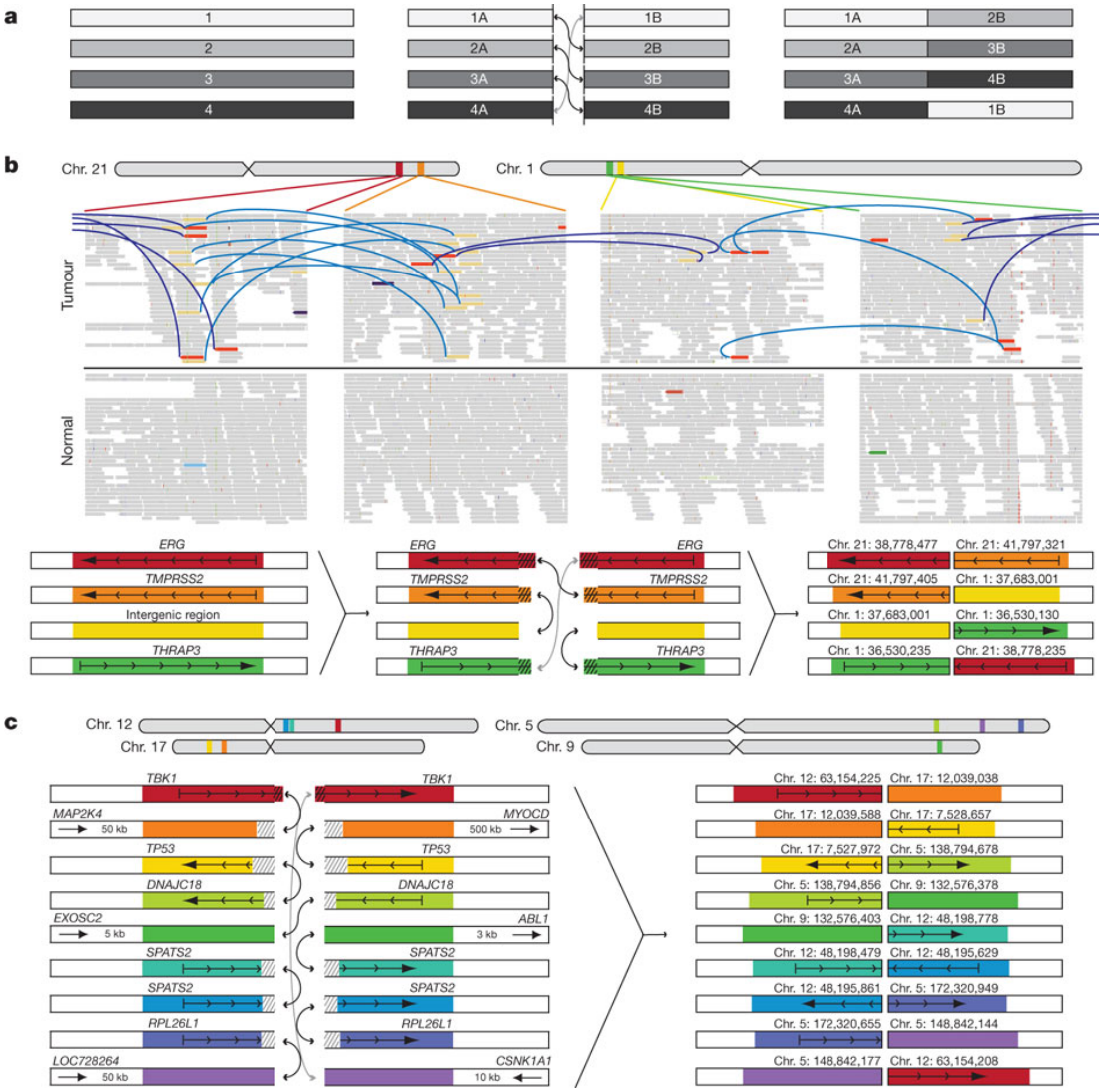


Figure 5-1 - **a. Diagram of 'closed chain' pattern of chromosomal breakage and rejoining.** Breaks are induced in a set of loci (left), followed by an exchange of free ends without loss of chromosomal material (middle), leading to the observed pattern of balanced (copy neutral) translocations involving a closed set of breakpoints (right). **b. Complex rearrangement in prostate tumor PR-1701.** TMPRSS2–ERG is produced by a closed quartet of balanced rearrangements involving 4 loci on chromosomes 1 and 21. Top, each rearrangement is supported by the presence of discordant read pairs in the tumor genome but not the normal genome (colored bars connected by blue lines). Thin bars represent sequence reads; directionality represents mapping orientation on the reference genome. Figures are based on the Integrative Genomics Viewer (http://www.broadinstitute.org/igv). Bottom, Diagram of breakpoints and balanced translocations. Hatching indicates sequences that are duplicated in the derived chromosomes at the resulting fusion junctions. **c. Complex rearrangement in prostate tumor PR-2832 involving breakpoints and fusions at 9 distinct genomic loci.** Hatching indicates sequences that are duplicated or deleted in the derived chromosomes at the resulting fusion junctions. For breakpoints in intergenic regions, the nearest gene in each direction is shown. In addition to the sheer number of regions involved, this complex rearrangement is notable for the abundance of breakpoints in or near cancer-related genes, such as TBK1, MAP2K4, TP53 and ABL1.

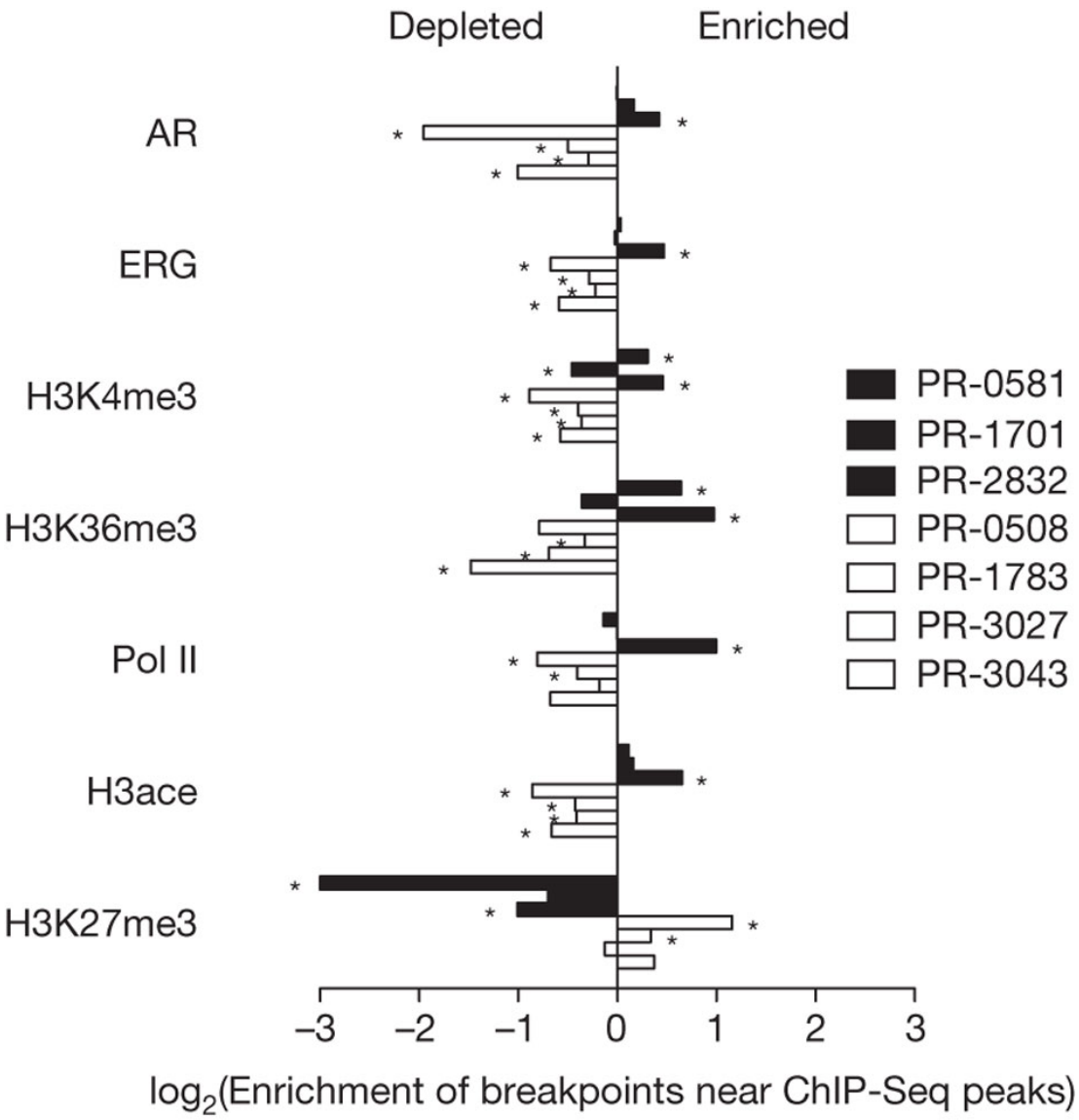


Figure 5-2 - **ChIP-seq binding peaks were defined previously for the TMPRSS2–ERG positive (ERG-positive) prostate cancer cell line VCaP**. For each genome, enrichment of breakpoints within 50Kbp of each set of binding peaks was determined relative to a coverage-matched simulated background (see Methods). TMPRSS2–ERG-positive prostate tumors are in black; ETS fusion-negative prostate tumors are in white. Enrichment is displayed as the ratio of the observed breakpoint rate to the background rate near each indicated set of ChIP-seq peaks. Rearrangements in ETS fusion-negative tumors are depleted near marks of active transcription (AR, ERG, H3K4me3, H3K36me3, Pol II and H3ace) and enriched near marks of closed chromatin (H3K27me3). P-values were calculated according to the binomial distribution. *Significant associations passing a false discovery rate cut-off of 5%.

Colorectal cancer rearrangements analysis

In colorectal cancer rearrangements play a major role as well. We analyzed nine colorectal tumors and identified 675 candidate somatic rearrangements. To assess the

accuracy of these findings, we tested 331 candidate somatic rearrangements by performing PCR across the putative junction in tumor and normal DNA; we pooled the PCR products and pyrosequenced them. We confirmed 92% of the calls as true somatic rearrangements; we found four calls (~1%) to be germline rearrangements and removed them from further analysis, and the remaining 22 calls (~7%) failed to yield PCR products in either tumor or germline DNA. Tumors with more somatic coding mutations also harbored more rearrangements ($R^2 = 0.55$).

Three samples showed clustering of inter-chromosomal translocations, where a series of rearrangements leads to extensive regional shuffling of two to three distinct chromosomes through balanced translocations. Because most of these events do not involve regions of substantial copy-number alterations, they represent a variant of chromothripsis involving alternating copy-number states induced by a single catastrophic complex genomic event. Our results show the potential for complex structural alterations to occur in regions of the genome that appear to be 'quiet' based on copy-number profiling.

Our discovery of recurrent VTI1A-TCF7L2 fusions is of particular interest. TCF7L2 encodes a transcription factor, known as TCF4 and belonging to the TCF/LEF family, that dimerizes with β-catenin (encoded by CTNNB1) to activate and repress transcription of genes essential for proliferation and differentiation of intestinal epithelial cells[72]. TCF7L2 is the most widely expressed member of the TCF/LEF family in colorectal cancer[73] and its expression is inversely associated with survival in colorectal cancer[74]. Moreover, the inherited risk of colorectal cancer is affected by polymorphisms in TCF7L2[75, 76] as well as by a polymorphism in an enhancer of MYC at which TCF4 and β-catenin cooperatively bind[77, 78]. Notably, TCF7L2 is known to harbor somatic point mutations in colorectal cancer[79, 80]. We additionally found a point mutation in CRC-5 affecting the splice-site at the 3′ end of exon 10, which is the exon encoding the HMG-box DNA binding domain, that would likely be a deleterious mutation.

Melanoma rearrangements analysis

We analyzed 25 metastatic melanomas and identified an average of 97 structural rearrangements per melanoma genome (range: 6–420). In addition to displaying a wide range of rearrangement frequencies, the proportion of intrachromosomal and interchromosomal rearrangements varied widely across genomes. Several

chromothripsis-like events have been detected. 106 genes harbored chromosomal rearrangements in two or more samples. Many recurrently rearranged loci contain large genes or reside at known or suspected fragile sites17; examples include FHIT (six tumors), MACROD2 (five tumors) and CSMD1 (four tumors). On the other hand, several known cancer genes were also recurrently rearranged, including the PTEN tumor suppressor (four tumors) and MAGI2 (three tumors), which encodes a protein known to bind and stabilize PTEN.

Multiple myeloma rearrangements analysis

We studied 38 multiple myeloma patients, performing whole-genome sequencing for 23 patients and whole-exome sequencing for 16 patients. We found 21 chromosomal rearrangements disrupting protein-coding regions, 4 of which affecting NF-κB pathway genes.

Discussion

Each study was the first whole genome sequencing analysis of its kind. Systematic genome characterization efforts have often focused primarily on gene-coding regions to identify 'driver' or 'druggable' alterations[79, 81, 82]. In contrast, the high prevalence of recurrent gene fusions has highlighted chromosomal rearrangements as critical initiating events in prostate cancer[83, 84]. Genome sequencing data indicate that complex rearrangements may enact pivotal gain- and loss-of-function driver events in primary prostate carcinogenesis. Moreover, many rearrangements may occur preferentially in genes that are spatially localized together with transcriptional or chromatin compartments, perhaps initiated by DNA strand breaks and erroneous repair. The complexity of 'closed chain' and other rearrangements suggests that complete genome sequencing-as opposed to approaches focused on exons or gene fusions-may be required to elaborate the spectrum of mechanisms directing prostate cancer genesis and progression.

In colorectal cancer, we have found the first recurrent fusion (VTI1A-TCF7L2). In prostate cancer, a positive correlation exists between the location of breakpoints in TMPRSS2-ERG-positive tumor cells and open chromatin in VCaP cells, and also between breakpoints present in ETS fusion-negative cells and VCaP regions of closed chromatin. This suggests that breakpoints may preferentially occur within regions of

open chromatin in some TMPRSS2-ERG-positive tumor cells while raising alternative possibilities for the genesis of breakpoints in ETS fusion-negative cells. Conceivably, somatic rearrangements may occur within regions of closed chromatin in tumor cells lacking ETS gene fusions. Alternately, such tumor cells may have distinct transcriptional or chromatin patterns, with many regions that are closed in VCaP being open in these cells. Clustering of breakpoints within active regions might also reflect selection for functionally consequential rearrangements during tumorigenesis. The relative contribution of these aspects to tumorigenesis will probably be informed by additional integrative analyses of epigenetic and structural genomic data sets across many tumor types.

Previous studies of genetically engineered mouse models have shown that the combination of ERG dysregulation and PTEN loss triggers the formation of aggressive prostate tumors[85, 86]. This same combination identifies a subtype of human prostate cancer characterized by poor prognosis[87]. The discovery of MAGI2 genomic rearrangements in prostate cancer suggests that interrogating both the PTEN and MAGI2 loci might improve prognostication and patient stratification for clinical trials of PI3 kinase pathway inhibitors. Additional mutated genes discovered in this study also suggest interesting therapeutic avenues. For example, the presence of point mutations involving chromatin modifying genes and the HSP-1 stress response complex (which includes the Hsp90 chaperone protein targeted by several drugs in development) raises the possibility that these cellular processes may represent targetable dependencies in some prostate tumors. Overall, complete genome sequencing of large numbers of relapsing primary and metastatic prostate cancers promises to define a genetic cartography that assists in tumor classification, elaborates mechanisms of carcinogenesis and identifies new targets for therapeutic intervention.

6. Non-linear quantification of pathway deregulation provides biologically relevant and compact representation of tumors

Publications #2 and #6.

Introduction

In recent years considerable effort was devoted to predict survival of patients using mRNA expression data, and many prognostic gene lists were offered, especially for breast cancer, where overtreatment is frequent due to lack of good predictors. Different studies, however, assembled very different gene lists, which generated considerable criticism[88-90]. As a response, some claimed that the different gene lists actually capture the same biological pathways and processes (e.g.[91]). I have tested this claim by comparing two celebrated prognostic signatures for breast cancer, I showed that the only biological process captured by both signatures was proliferation, a process whose relevance and prognostic value was well known, and clinically measured, long before gene expression profiling. This might suggest that most current expression based prognostic signatures do not capture the relevant biology, and different approaches are required. This effort has been published[92] (publication 6 of the list presented in chapter 9) and appears here as Appendix A.

The lack of current ability to use existing machine learning tools to predict survival from mRNA expression data inspired me to develop a general method that will process these data to deduce comprehensive biologically relevant information, using existing biological knowledge. Established gene-level analysis methods can then be used on the processed information to gain a more biologically relevant outcome. This work, described below, together with an analysis of Pathifier results for colorectal cancer (as analyzed by Dr. Michal Sheffer), has been summarized in a manuscript currently under revision (publication 2 in chapter 9).

The operation of many important pathways is altered during cancer progression, to allow the tumor to prosper, proliferate and overcome the cellular defense systems. Hence, identifying the involved pathways and characterizing when and to what extent they are disrupted was recognized as a promising way to attain improved understanding of different cancers (e.g.[93-95]). Moreover, advanced therapies often target a specific pathway, and hence both the development and clinical

implementation of effective personalized cancer treatments will be significantly facilitated by such a pathway-level understanding. In recent years many large datasets, mostly of mRNA expression, have become available, creating an opportunity to infer such important pathway level information. Indeed, much effort has been invested in methods for pathway analysis of data from cancer, as well as for other diseases (reviewed in[95-103]). However, almost all available methods focus on dataset level analysis, that is, they don't use gene-set-based characterization of a specific sample. In parallel with my efforts, Vaske et al. published PARADIGM[104], a method which deduces pathway activity from mRNA expression and DNA copy number data, for each individual sample, on the basis of known pathway structure. PARDAGIM is a powerful tool that allows deducing the level of a known interaction or a simple downstream activity. However, since for such analysis one needs detailed insight into the networks and mechanism of pathway activity, which is not always available, as well as necessary relevant data such as protein abundance and phosphorylation, PARDAGIM is not best suited for analysis of complex pathways in large cancer data sets at this point. In my work I offer a different approach for pathway-based sample characterization, which does not depend on existing detailed reliable knowledge of the network underlying pathway activity, and can estimate pathway abnormality or disruption in a given sample, and not just downstream activity levels of simple pathways. I show that this representation is indeed clinically relevant, and complementary to PARADIGM.

Results

Brief outline of the Method (see Methods for more details)

Pathifier analyzes N_P pathways, one at a time, and assigns to each sample i and pathway P a score $D_P(i)$, which estimates the extent to which the behavior of pathway P deviates, in sample *i,* from normal. To determine this *pathway deregulation score* (PDS), I use the expression levels of those d_P genes that belong to P (e.g. using databases such as[105-108]). Each sample i is a point in this d_P dimensional space; the entire set of samples forms a cloud of points, and I calculate the "principal curve"[109], that captures the variation of this cloud. Next I project each sample onto this curve; the PDS is defined as the distance $D_P(i)$, measured along the curve, of the projection of sample i, from the projection of the normal samples (see Figure 6-1 below). On the

basis of genome-wide gene-level expression data I generate a pathway-level, biologically relevant N_P-dimensional representation of each sample, and mine this representation for insights. *Pathifier* allows integration of external (partial and approximate) information on the samples, such as tumor stage, grade, chromosomal instability, survival, etc.

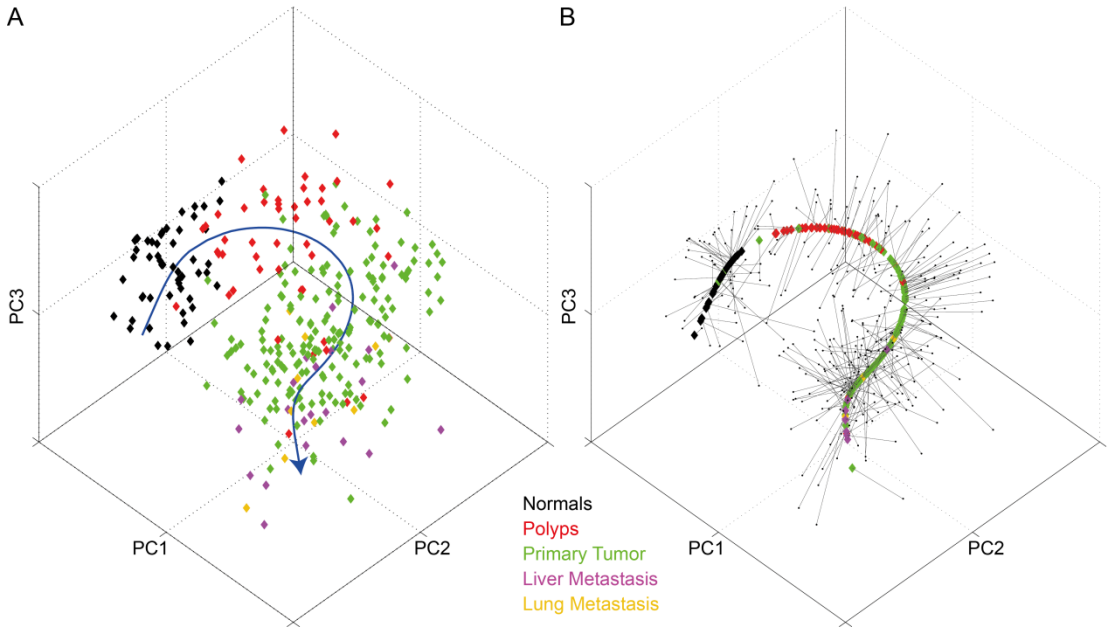


Figure 6-1 - **The principal curve learned for the apoptosis KEGG pathway on the Sheffer et al.[110] colorectal dataset.** The data points and curve are projected onto the three leading principal components. **A.** The principal curve (in blue) going through the cloud of samples. **B.** The samples projected onto the curve.

<u>High values of inferred pathway deregulation scores correspond to activating mutations in glioblastoma</u>

Pathifier was first applied to glioblastoma multiforme (GBM) gene expression measurements of 445 tumors and 10 adjacent normal brain tissues collected by The Cancer Genome Atlas (TCGA)[82]. I successfully estimated deregulation scores for 548 pathways from KEGG[105, 106], BioCarta[107] and the NCI-Nature Pathway Interaction Database[108] for the analysis. The results of the analysis are summarized in a 548 by 455 table, representing the deregulation score of each pathway in every tumor (shown in Fig 6-2A). Out of the 445 samples, 135 were sequenced, to detect point mutations in key genes. Ten genes (*EGFR, PIK3R1, IDH1, ERBB2, DST, SYNE1, NF1, RB1, PTEN, TP53*) were mutated in more than 5% (7 or more) of the samples. In 96 of the 135 samples one or more of these 10 genes had mutations. 94 pathways are significantly related to one or more of the mutations (Mann–Whitney, FDR<1%).

Hierarchical average linkage clustering of the pathway deregulation scores of these 94 pathways reveals 3 pathway clusters (Fig. 6-2C & Supplementary Table 1 in Appendix B). Pathways of cluster P1 are deregulated mostly in samples of cluster S2, which comprises tumors with IDH1 mutation. All 32 pathways in P2 are activated by *EGF* signaling. Indeed, they are highly deregulated on sample cluster S5, which includes almost all patients with *EGFR* mutations. The fact that the PDS capture the deregulation of EGF signaling pathways, expected in samples with oncogenic EGF mutations[111], is reassuring and indicates that Pathifier indeed captures relevant biological information. Another example of concordance of PDS with mutations is of cluster P3, which contains many pathways with high PDS in tumors with NF1 mutations (mostly in sample cluster S4), and low PDS in tumors with IDH1 mutation (mostly in S2).

The scores of many cell death-related pathways are correlated with necrosis levels of glioblastoma

The PDS of 242 pathways significantly correlate with the necrosis levels of the samples, as quantified by TCGA (FDR<1%), see Supplementary Table 2 in Appendix B. Some of these pathways are indeed expected to cause cell death, such as: *SODD* signaling, *FAS* pathway, *NEF* pathway, *BAD* phosphorylation pathway, apoptosis, caspase pathway, Notch signaling, Induction of apoptosis through DR3 and DR4/5 Death Receptors, p75(NTR)-mediated signaling, oxidative stress induced gene expression via *NRF2* and *ERK5* in neuronal survival. Many of the other pathways are growth factor pathways, such as: *NGF*, *ERBBs*, *PDGFRB*, IGF, and Trk receptor pathways. A few hypoxia and angiogenesis related pathways are also correlated with necrosis (*VEGF* pathways, HIF pathways, angiopoietin receptor pathway, lymphangiogenesis pathway, Hypoxia and p53 in the Cardiovascular system).

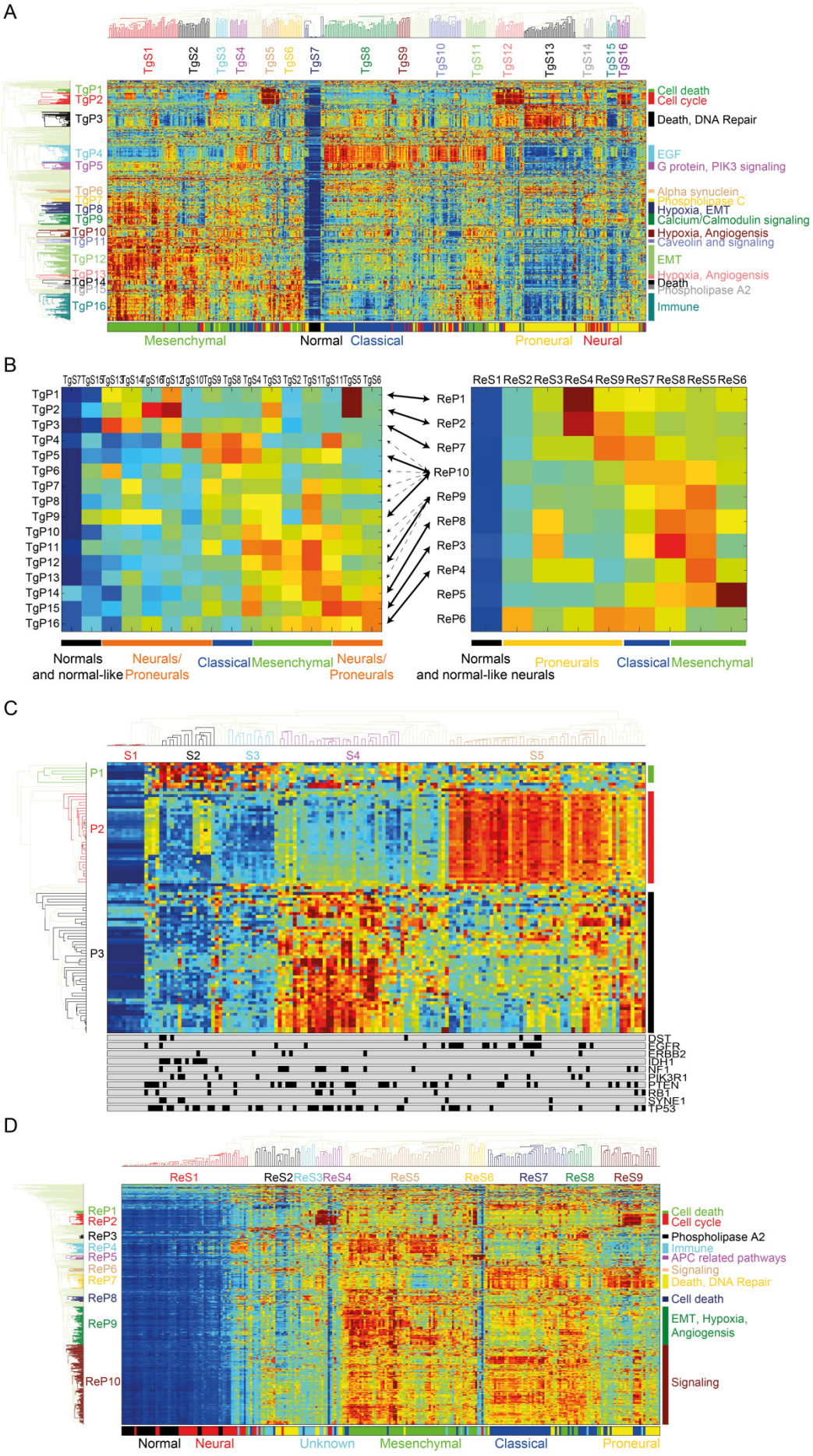

A
TgS1 TgS2 TgS3 TgS4 TgS5 TgS6 TgS7 TgS8 TgS9 TgS10 TgS11 TgS12 TgS13 TgS14 TgS15 TgS16
TgP1 TgP2 TgP3 TgP4 TgP5 TgP6 TgP7 TgP8 TgP9 TgP10 TgP11 TgP12 TgP13 TgP14 TgP15 TgP16
Cell death
Cell cycle
Death, DNA Repair
EGF
G protein, PIK3 signaling
Alpha synuclein
Phospholipase C
Hypoxia, EMT
Calcium/Calmodulin signaling
Hypoxia, Angiogensis
Caveolin and signaling
EMT
Hypoxia, Angiogensis
Death
Phospholipase A2
Immune
Mesenchymal
Normal
Classical
Proneural
Neural
B
TgS7 TgS15 TgS13 TgS14 TgS16 TgS12 TgS10 TgS9 TgS8 TgS4 TgS3 TgS2 TgS1 TgS11 TgS5 TgS6
TgP1 TgP2 TgP3 TgP4 TgP5 TgP6 TgP7 TgP8 TgP9 TgP10 TgP11 TgP12 TgP13 TgP14 TgP15 TgP16
ReP1 ReP2 ReP7 ReP10 ReP9 ReP8 ReP3 ReP4 ReP5 ReP6
ReS1 ReS2 ReS3 ReS4 ReS9 ReS7 ReS8 ReS5 ReS6
Normals and normal-like
Neurals/ Proneurals
Classical
Mesenchymal
Neurals/ Proneurals
Normals and normal-like neurals
Proneurals
Classical
Mesenchymal
C
S1 S2 S3 S4 S5
P1 P2 P3
DST
EGFR
ERBB2
IDH1
NF1
PIK3R1
PTEN
RB1
SYNE1
TP53
D
ReS1 ReS2 ReS3 ReS4 ReS5 ReS6 ReS7 ReS8 ReS9
ReP1 ReP2 ReP3 ReP4 ReP5 ReP6 ReP7 ReP8 ReP9 ReP10
Cell death
Cell cycle
Phospholipase A2
Immune
APC related pathways
Signaling
Death, DNA Repair
Cell death
EMT, Hypoxia, Angiogensis
Signaling
Normal
Neural
Unknown
Mesenchymal
Classical
Proneural

Figure 6-2 - **A. Clustering of all pathways of the TCGA GBM dataset.** Each row corresponds to a pathway and each column – to a sample. Pathways and samples are clustered according to pathway scores. Blue color represents low score ("no deregulation"), and red high. The bottom bar represents the glioblastoma subtype. Notice that the pathway based clustering captures the subtypes well, and identifies a secondary sub-stratification. **B. Summary of clustered pathway scorers for the TCGA (left) and REMBRANDT (right) glioblastoma datasets.** Each row corresponds to a pathway cluster and each column to a sample cluster, displaying the median value of deregulation for each pair of clusters. Arrows connect between pathway clusters that match (that is, the pathways in the clusters significant overlap). When few significant matches are possible (for ReP9 and ReP10) all are shown in dashed arrows, except for the extreme significant ones ($p<10^{-5}$). Some of the Neurals/Proneurals are mostly not deregulated, and some are deregulated on TgP1/TgP2/TgP3 or matching ReP1/ReP2/ReP7. Classical tumors are deregulated on TgP4/TgP5 and possibly TgP6/TgP7 as well as matching ReP10 (and unmatched ReP6/ReP7). Mesenchymal samples are highly deregulated on TgP8-TgP16 as well as matching ReP10/ReP9/ReP8/ReP3/ReP4 (and unmatchable ReP5). The Classical-Mesenchymal cluster TgS4 matches ReS8, and indeed they are both deregulated on the matching TgP4/TgP5/TgP10/TgP11/TgP12/TgP14/TgP15 and matching ReP10/ReP9/ReP8/ReP3 (as well as unmatchable ReP5). **C. Deregulation scores of 94 pathways correlated with mutations.** The bottom bars display the mutation status, each bar for one gene (samples with mutation are marked green). Cluster S1 corresponds to normal samples, S2 mostly to samples with IDH1 mutations, S4 mostly to samples with NF1 mutations, and S5 mostly to samples with EGFR mutations. Notice pathway cluster P2, which consist mostly of EGF activated pathways, and is highly deregulated on the EGFR mutated samples. **D. Pathway deregulation scores for the REMBRANDT GBM dataset.** As in panel A, the pathway clusters correspond to the subtypes but offer additional sub-stratification. See panel B for a concise representation.

PDS-based stratification of glioblastoma

Hierarchical average-linkage clustering according to PDS of the TCGA data (Fig. 6-2A) generates (a) sample clusters, which are consistent with known classification and extend it, and (b) pathway clusters, with related biological functions. The Normal samples form cluster TgS7; Mesenchymal tumors comprise clusters TgS1-3 (and the small cluster TgS11); Classical cancers are in TgS8-9, while Neurals and Proneurals in TgS12-16. A concise representation of the characteristic deregulation profiles of the sample types over the pathway clusters is given on the left side of Figure 6-2B.

Pathway cluster TgP1 consists of cell cycle arrest and cell death pathways; TgP2 contains cell cycle pathways and many of KEGG's "cancer" pathways (including glioma) which capture cancer progression and signaling; TgP3 contains mainly cell death and DNA repair pathways and is deregulated mostly on the Neural and Proneural samples; The pathways of cluster TgP4 correspond to the EGF activated

pathways mentioned above; Cluster TgP5 contains pathways that are deregulated mostly on the Classical samples. Some of them are indeed suspected to be specific to this subtype, such as hedgehog-GLI signaling[112] and GPCR/CXCR4 signaling[113] while the deregulation of some others in this subtype is a new prediction: such as PAR1(*F2R*)-mediated thrombin signaling, axon guidance, etc.; Half of the TgP6 pathways involve alpha synuclein amyloids; All TgP7 pathways involve phospholipase C; the pathways that comprise clusters TgP8-TgP10, TgP12-TgP15 belong to the 242 pathways that were correlated with necrosis that were mentioned above, and are also highly expressed on many Mesenchymal samples. As mentioned, many of these pathways (and specifically the pathways of TgP8-TgP10, TgP13 and TgP15) are related to hypoxia and angiogenesis, and I find, in agreement with previous knowledge, that they score higher in Mesenchymal glioblastoma[114]. Clusters TgP8 and TgP12 contain several Epithelial-Mesenchymal Transition(EMT) related pathways (such as N-cadherin signaling, epithelial tight junctions, Rho/Rac/CDC42 signaling, regulation of actin cytoskeleton, ECM-receptor, Focal adhesion) obviously related to Mesenchymal tumors; 7 of the 8 pathways of TgP11 are key signaling pathways involving caveolin; TgP12 contains many of the pathways correlated with NF1 mutation (P3 above); TgP14 contains mostly cell death pathways; TgP15 pathways all involve phospholipase A2; TgP16 contains many immune pathways.

Notably, the pathway score based clustering suggests also further, more subtle, stratification of the tumors. Most of the Mesenchymal samples are in clusters TgS1, TgS2, TgS3 and TgS11, mostly deregulated in pathway clusters TgP8-TgP16. The main differences between the sub-types are the lack of deregulation of TgP7-TgP9 pathways in the samples of TgS2; the deregulation of TgP2 in TgS3 and of TgP4 pathways in TgS11. TgS11 and more so TgS4 contain Classical-like Mesenchymals and Mesenchymal-like Classicals – these intermediate tumor types are deregulated on *both* the typical Mesenchymal pathway clusters TgP8-TgP15 and the characteristic Classical TgP4 and TgP5. The emergence of this "sub-class" might be due to heterogeneous samples containing both types of cells, or to a genuine new subtype with Classical *and* Mesenchymal features. The Neural and Proneural samples are divided into 9 clusters. The bulk of the samples are in clusters TgS12-TgS16. The normal-like Neural/Proneural tumors of TgS13 and TgS15 are not deregulated on most pathways.

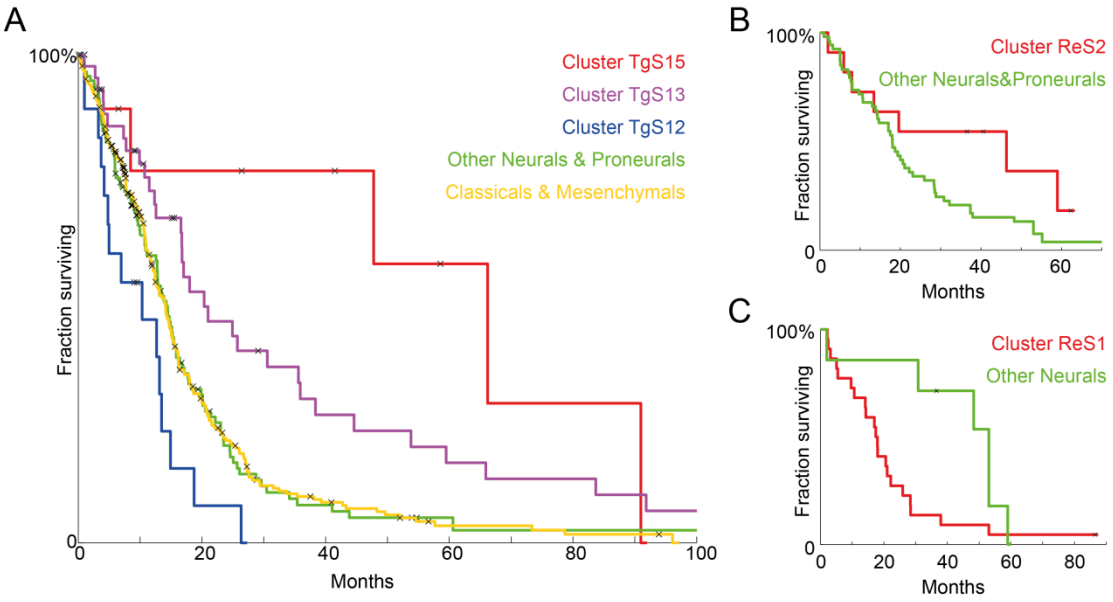


Figure 6-3 - **Kaplan Meier plots of Neural and Proneural sub-stratification. A.** In the TCGA dataset, patients in clusters TgS13 and TgS15 have better prognosis. Neural and Proneural tumors were divided to three groups – those in cluster TgS12 (in blue), those in TgS13 (in purple), those in TgS15 (in red) and all others (in green). Kaplan-Meier plots show clear separation between the four, where cluster TgS15 patients survive the longest (p=0.009) and cluster TgS13 little less, but still better compared to the others (p=0.015), and those in TgS12 significantly survive less than the others (p=0.003). The prognosis of the neural and proneural tumors that are not in TgS12, TgS13 or TgS15 (in green) is similar to Classical and Mesenchymal tumors (in yellow). **B.** In the REMBRANDT dataset, ReS2 patients have better survival. Neural and Proneural tumors were divided into two groups – those in cluster ReS2 (in red), and all others (in green). Kaplan-Meier plots show clear benefit for the ReS2 patients (p p=0.066). **C.** In the REMBRANDT dataset, ReS1 patients survive less. Cluster ReS1 contains only normal samples and normal-like Neural samples. Interestingly, these Neural tumors (in red) have significantly worse prognosis (p=0.032) than other Neurals (in green).

To validate these results I used data from REMBRANDT (REpository for Molecular BRAin Neoplasia DaTa)[115, 116]: the clustering results are shown in Figure 6-2D. I used Fisher's exact text ($p<0.05$) to test for correspondence between the pathway clusters found in the TCGA data and the ones in REMBRANDT data (marked ReP). Cluster ReP1 matches TgP1, ReP2 matches TgP2, ReP3 matches TgP15, ReP4 matches TgP16, ReP7 matches TgP3, ReP8 matches TgP14, ReP9 includes parts of TgP10-TgP13 (most strongly related to TgP12), and ReP10 includes parts of TgP4-TgP9 (most strongly related to TgP5 and TgP9). Under this mapping, indeed similar characteristics emerge (Fig. 6-2B). Some of the Neurals/Proneurals are mostly not deregulated (ReS1/ReS2 vs. TgS7/TgS15/TgS13) and some are deregulated on TgP1/TgP2/TgP3 or the matching ReP1/ReP2/ReP7. Classical tumors are deregulated

on TgP4/TgP5 and possibly TgP6/TgP7 as well as on the matching ReP10 (and unmatched ReP6/ReP7). Pathways of clusters TgP8-TgP16 as well as the matching ReP10/ReP9/ReP8/ReP3/ReP4 (and unmatchable ReP5) are highly deregulated in the Mesenchymal samples. The Classical-Mesenchymal cluster TgS4 matches ReS8, and indeed the corresponding samples are deregulated on TgP4-TgP5/TgP10-TgP12/TgP14-TgP15 and, respectively, on the matching ReP10/ReP9/ReP8/ReP3 (as well as on the unmatchable ReP5).

The pathway based sub-stratification of GBM has important clinical implications

Neural and Proneural samples are thought to have better prognosis[112, 114]; the pathway based sub-stratification reveals, however, that this notion is due to a subset of better survivors (logrank p-value[117]<0.05). In the TCGA data, patients of clusters TgS15 and TgS13, which have relatively few deregulated pathways, survive significantly longer than other Neural and Proneural samples (p=0.009 for TgS15 and p=0.015 for TgS13), while patients from TgS12 have worse prognosis (p=0.003) as can be seen in Figure 6-3A. If this group of good survivors is removed from the Neural and Proneural samples, the remaining patients of these classes do no better than patients with Mesenchymal and Classical tumors. The separation between survival of the patients of TgS12,13,15 remains significant even if the comparison is made only for the Proneural samples.

These results are reproduced on the REMBRANDT data as well – patients with Neural or Proneural tumors in cluster ReS2, the one for which only few pathways are deregulated, have better prognosis than other Neurals and Proneurals (p=0.066, Fig. 3B). Cluster ReS1 contains only normals and normal-like Neural samples. Interestingly, these normal-like Neural tumors have worse prognosis than other Neurals (p=0.032, Fig. 6-3C).

Pathways associated with survival in glioblastoma

77 pathways are significantly related to survival on the TCGA data, and 187 on the REMBRANDT data (FDR < 10%, from Kaplan-Meier analysis, comparing the top 1/3 deregulated samples to the bottom 1/3, logrank p-value). 37 of these pathways overlap, constituting a significant intersection (p=0.005). For all but two pathways, higher disruption scores were correlated with bad prognosis on both datasets. These two pathways are probably false positives, as higher deregulation implied worse

prognosis on the REMBRANDT dataset but better prognosis on the TCGA dataset. The other 35 pathways are listed in Table 1.

Many of the pathways found to be related to patient survival make biological sense: Some are related to angiogenesis, critical to glioblastoma progression (such as VEGF signaling, Fibrinolysis, PDGFRβ signaling, α4β1 integrin signaling, HIF2α pathway); Many are known key players in glioblastoma and cancer in general, are known to have a prognostic value, and are promising drug targets (such as MAP kinase[118-122], Insulin signaling and its components[123-125], RET tyrosine kinase[126], EGFR/ERBB signaling[127], PDGF signaling[128] and integrins[129]); Agrin deregulation may temper the blood-brain barrier in glioblastoma[130]; Growth hormone (GH) plays a crucial role in stimulating and controlling the growth, metabolism and differentiation of many mammalian cells, and hence clearly relevant for cancer aggressiveness[131]; The hematopoiesis pathway contains cytokines and it is suspected to be related to cancer progression and drug resistance by interactions with the immune system[132-134]; Linolenic acids and their products were suggested to prolong cancer patient survival[135]; FcεRI may protect against cancer by IgE antitumor immunity[136]; Cell-matrix adhesions are clearly related to invasion and metastasis[137]; *GnRH* is a neurohormone that may drive proliferation in glioblastoma and other cancers, and therefore is also a suggested drug target[138-143]; Phosphatidyl-inositol 3- and 4-kinases, key ingredients of the inositol phosphate pathway, are known to have important roles in glioblastoma and cancer in general, and hence are possible drug targets[144-146]; Surprisingly, cholera toxin was also found related to glioblastoma; WNT signaling has a key role in brain and other cancers, and is related to cancer stem-like cells and poor prognosis[147-149]; Alterations in E-cadherin mediated cell-cell adhesion are associated with an increase in carcinoma cell motility, invasiveness and metastasis[150]; Glypican-1 is crucial for efficient growth, metastasis, and angiogenesis of cancer, and lack of it slows down pancreatic tumor progression[151]; Fas and TNF-alpha are key players in apoptosis whose deregulation is a clear hallmark of cancer[152153]; α4β1 integrin is related to angiogenesis[154], and is involved the survival and chemoresistance of several types of cancer[155]; β2 integrins are known to predict poor survival in blood cancers[156, 157], and may cause fibrinolysis via uPA/uPAR; α6 integrin regulated stemness and invasiveness in glioblastoma[158, 159], and has a prognostic value in breast cancer[160]; P53 is a key player in glioblastoma and cancer in general; *PTPN1* (aka PTP1B) may promote apoptosis in cancer and is a possible drug target for gliomas[161];

Reelin is a secreted glycoprotein guiding migratory neurons, it is downregulated in neuroblastoma, which may contribute to metastasis[162, 163]; Syndecans induce proliferation and invasion[164-166] , and may serve as a prognostic predictor[167, 168].

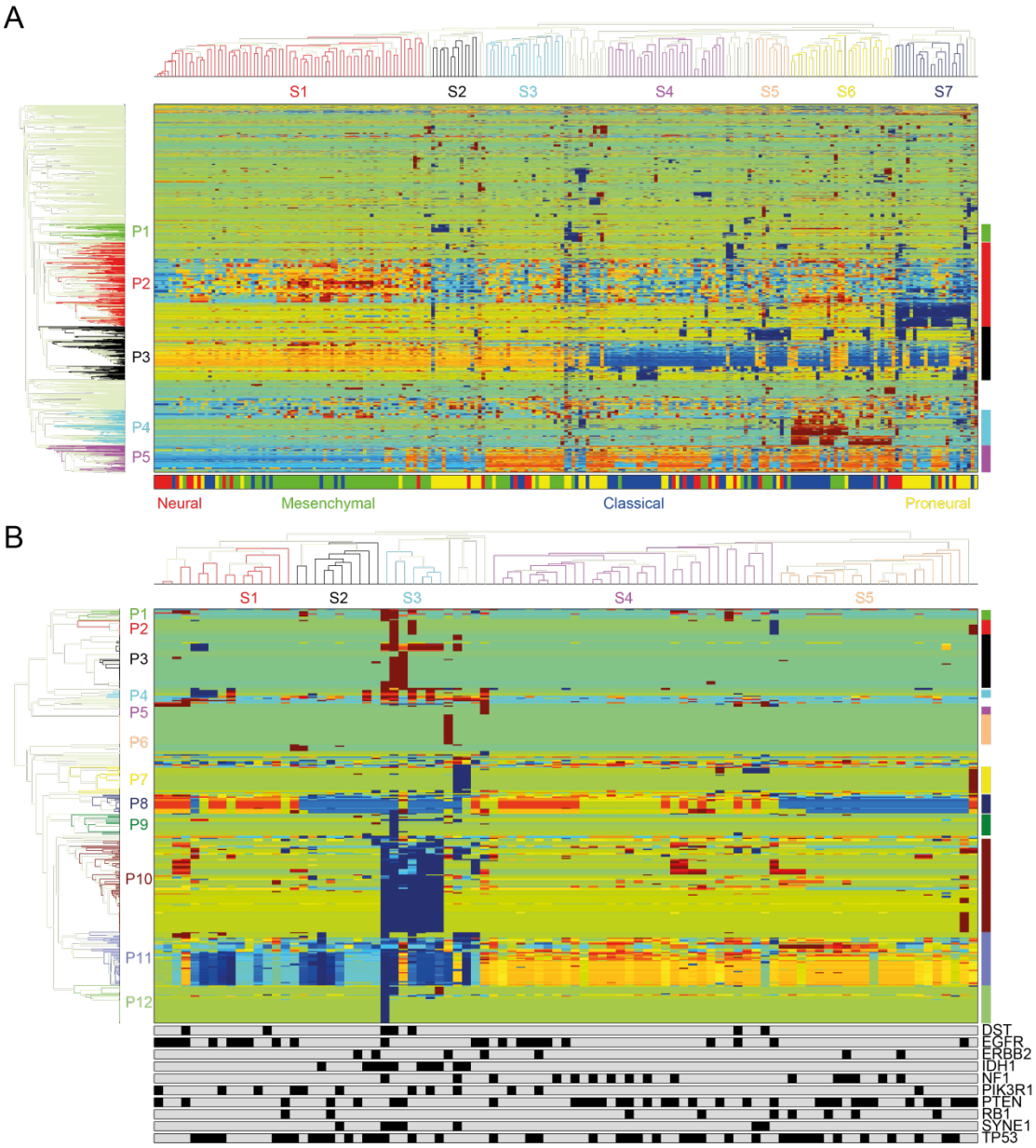


Figure 6-4 - **A. PARADIGM's IPAs for the TCGA GBM dataset.** Each row corresponds to a PARADIGM "entity" and each column to a sample. Entities (pathways, interactions, complexes etc.) and samples are clustered according to the IPAs. Blue color represents low activity, and red high activity. The bottom bar displays the subtype. **B. PARADIGM's IPAs correlated with mutations.** The bottom bars display the mutation status for the corresponding gene.

Comparison with PARADIGM

PARADIGM and Pathifier are complementary approaches. I have repeated the analyses using the scores derived using PARADIGM. *EGFR* gene and *EGFR* complexes were indeed found to be correlated to *EGFR* mutations as expected (FDR<1%). However, no pathways were found to be correlated to the *EGFR* mutations. Not much more could be inferred by the analysis based on PARADIGM scores analysis (Fig. 6-4 above). As reported in[104], some stratification of glioblastoma is possible by PARADIGM IPA's, but it does not match strongly the known subtypes. Though it did successfully detect a relevant cluster of *HIF1A* low and E2F high tumors, PARADIGM missed many of the observations I mention above. None of the IPA's is found to be related to survival with FDR of 19% or less. Therefore I conclude that while PARADIGM is a very useful method to integrate different types of data and deduce simple complexes and downstream activations (such as EGF receptor activation), Pathifier provides additional clinically relevant and easy to interpret information about the deregulation of complex pathways.

Methods

Scoring a gene set

Denote by S_P the d_P-dimensional space, where each coordinate is the expression level of a gene that belongs to a given pathway P, and represent each sample by a point in this space. Pathifier looks for a one-dimensional curve in S_P (or in a subspace of S_P) that best describes the variability (e.g., due to disease progression) of the samples across S_P. That is, a curve that passes through the "middle of the cloud" of samples, assuming that any two points (samples) that have proximal projections onto the curve, also share similar pathway functionality.

Variance stabilization: Since for some genes large variation of expression is observed, while for others a similar effect on a pathway's functionality may be induced by a smaller variation, Pathifier doesn't use the absolute expression values. Rather, a gene's expression values are divided by the standard deviation of its expression in some normalization set of samples (such as normal samples, or the entire set of samples). To avoid genes whose entire variation is due to noise, Pathifier omits from the outset genes with little variation over all samples.

Correlations: Many of the genes in the gene set of a pathway might be highly correlated, conveying the same information, while some other important information might reside in a single gene in the set. To counter this effect, and to improve the running time, Pathifier doesn't actually search for a curve in S_P, but in a space S_P' of smaller dimensionality k, identified as follows. First it performs PCA and picks only the principal components with high variance (I chose 1.1 as a threshold, i.e. keeping PC along which the variance exceeds by more than 10% that of the normalization set). The number of such PCs is k and they define the space S_P', in which the entire ensuing computation is done.

Principal curve: I use Hastie and Stuetzle's algorithm[109] to find a principal curve in S_P' (see Fig. 1). For its semi-supervised variant I modify the method to allow incorporation of available independent partial knowledge on the variability (see below). After such a curve is found, Pathifier projects each point x_i, that represents sample i in S_P', onto f_i, its closest point on the curve. The deregulation score $D_P(i)$ of sample i is defined as the distance along the curve between f_i and a reference point r. If additional (supervised) ranking is provided, the reference point r is chosen to be the projection of the sample with the lowest rank. If no guiding ranking is provided, I define the reference point r as the centroid of some reference set of samples. In this case the reference set is used to define the curve's direction, by making sure the point representing the median coordinates of the reference set is closer to the beginning of the curve (otherwise flip the curves direction). In this study, the reference set is comprised of the healthy samples from the same tissue (henceforth "normal samples"), which indeed tend to concentrate on one side of the curve, due to the high similarity among normal samples and the large difference from tumor samples. The distance $D_P(i)$ provides a measure of the extent to which the expression levels of the genes associated with pathway P were perturbed in sample i by the disease.

In some cases the normal samples fall roughly in the middle of the curve. When this happens, the curve captures two different kinds of deregulation, with samples moving away from the normal samples along two distinct paths. In principle one can use other (than normal) samples as reference, though doing this makes sense only in cases when the inner variability of the new reference set is considerably smaller than the overall variability.

Finding a stable gene set

Often some of the genes in the gene set are noisy (in the sense that their variation does not reflect information relevant to the biology I'm trying to capture), and I would rather omit them. Since Pathifier work in S_P' and not in S_P, it actually omits metagenes (linear combinations of genes), but similar considerations imply that some of the metagenes might be noisy and should be omitted. This is partly taken care of by omitting genes and metagenes that don't vary much, but some of the noise might be due to highly varying metagenes, where most of the variation is unrelated to the biological information captured by the gene set.

To find out which metagenes should be omitted, Pathifier selects, one at a time, those along which the samples were farthest from the curve, as expected for noisy metagenes, and finds after each omission the new corresponding principal curve. To assess which curve is the best, Pathifier checks the sensitivity of the gene set's scores to sampling noise (the variance over randomly selecting, 100 times, 80% of the samples). If there is a significant improvement in the stability, Pathifier omits the metagene whose omission yielded the most stable curve, and continues in a greedy fashion. If the improvement is not significant (or stability actually becomes worse), it stops.

Principal Curves

The concept of principal curve was first proposed by Hastie and Stuetzle[109] as a non-parametric nonlinear extension of the linear Principal Component Analysis. Denote by *f(λ)* a curve in *p*-dimensional space, where λ is a single parameter whose variation traces all the points along the curve. A curve *f* is defined to be a principal curve *associated with a distribution P(x)* defined over some space, if and only if it is a smooth, one dimensional non intersecting curve that is *self-consistent*, i.e. each point *y* on the curve is the expected value of all the points *x* whose projection onto the curve is *y*. Let the projection index $\lambda_f(x)$ be the value of λ for which the projection of *x* on the curve is *f(λ)*, i.e. :

$$\lambda_f(x) = \sup_{\lambda}\left\{\lambda \middle| \|x - f(\lambda)\| = \inf_{\mu}\|x - f(\mu)\|\right\}$$

The condition for self-consistency is simply

$$f(\lambda) = \mathbb{E}\left(x \middle| \lambda_f(x) = \lambda\right)$$

Since in practice Pathifier gets a finite data set *X,* of *n* points in d_P dimensional space $X \in M_{n \times d_P}(\mathbf{R})$, while the distribution it is sampled from is not known, scatterplot smoothing is used. Hastie and Stuetzle also offer a two-steps iterative algorithm for finding such a principal curve:

1. Conditional-Expectation step: Fix $\lambda = \{\lambda_i\}_{i=1}^n$ and minimize $\mathbb{E}\|X - f(\lambda)\|$ by setting $f(\lambda_i)$ to be the local average (of the points projected onto a neighborhood of $f(\lambda_i)$, e.g. the points x_j for which $\lambda_j \cong \lambda_i$).
2. Projection step: Given $f = \{f(\lambda_i)\}_{i=1}^n$ find for each x_i the corresponding value of $\lambda_i = \lambda_f(x_i)$ assuming f is piecewise linear.

The line along the first linear principal component is used as a starting curve, and the algorithm is iterated until convergence.

Since Hastie and Stuetzle's seminal publication, many alternatives and generalizations were offered[169-174], but they all focused on unsupervised learning, not allowing the incorporation of additional information. In particular, another non-linear trajectory, defined in the full expression space, was used to capture tumor progression[175]. I offer a modification to Hastie and Stuetzle's algorithm that allows introduction of prior belief on the order of the curve parameter $\{\lambda_i\}_{i=1}^n$. This prior affects the estimation of the distribution from which *X* is sampled, and therefore generates a modified principal curve.

Data and preprocessing

GBM mRNA expression data was downloaded from TCGA data portal on April 2011[82]. To reduce batch effects of arrays and protocols, I used only Agilent G4502A arrays measured at the UNC medical school, yielding 455 samples, 10 of which were from normal tissue. Additionally, 228 glioblastoma samples and 28 normal brain samples were obtained from REMBRANDT[116]. Subtypes classification for 197 TCGA samples was taken from Veerhak et al.[112] Classification of the REMBRANDT data and additional TCGA samples was done using the same genes. For TCGA data we used level 3 already processed data and for REMBRANDT data was summarized with PLIER and normalized with cyclic LOWESS[176]. To eliminate noisy genes only the 5000 most varying genes for each cancer type (sum of variation on the two datasets) were selected for further analysis.

Assembly of pathway associated gene sets

Gene sets were imported from three pathway databases, KEGG[105, 106], BioCarta[107] (both downloaded from MSigDB[177]) and the NCI-Nature curated Pathway Interaction Database[108]. Identity of genes in gene sets was decided according to their official gene symbols. Gene sets with less than 3 genes varying in the data were omitted, leaving 173 KEGG pathways, 188 BioCarta pathways and 197 NCI-Nature PID pathways.

Integration of External Information: Semi Supervised Principal Curve

Finding a principal curve is closely related to finding the "correct" order to go through the data points. A curve explicitly defines that order (the ranking of $\{\lambda_i\}_{i=1}^n$), and given an order, a curve could be deduced, e.g. by Hastie and Stuetzle's[109] Conditional-Expectation step (or any other data fitting approach). Therefore, if one has detailed information on the order, finding the curve is very simple. However, if there is only vague or imprecise information, one faces the full problem of finding a principal curve, but the information still might help improve the identification of the desired curve. I capture this knowledge by assuming that the Spearman correlation of $\lambda = \{\lambda_i\}_{i=1}^n$ and a vector of some (approximately known) ranking is high. I have chosen Spearman correlation since (a) it requires only the knowledge of ranks (allowing ties to represent uncertainties), and not the explicit values of some variable, and (b) it does not depend on the specific details of the curve parameterization.

Notice that in Hastie and Stuetzle's algorithm, the conditional-expectation step is not affected by the ranking prior, as the $\{\lambda_i\}_{i=1}^n$ are given and fixed. Therefore I adapt only the projection step, in a way that incorporates the prior. The projection step is equivalent to minimizing $\left\|x_i - f(\lambda_i)\right\|$ for every $1 \le i \le n$. I'm interested in simultaneously maximizing the Spearman correlation $\rho = 1 - \frac{6\sum d_i(\lambda_i)^2}{n(n^2-1)}$, where $d_i(\lambda_i)$ is the difference between the rank of λ_i and the rank of i in the given guiding ranks vector. Therefore, the minimized function can be written as $(1-\alpha)\sum_{i=1}^{n}\left\|x_i - f(\lambda_i)\right\|^2 + \alpha\frac{6\sum d_i(\lambda_i)^2}{n(n^2-1)}$, where α reflects the confidence in the ordering.

Thanks to the iterative framework, it is not necessary to find the global minimum in each step. Therefore, I take a greedy approach to save computation. First Pathifier computes for every x_i the distance to each segment of the curve, and then it selects the $\nu_i = \lambda_i$ that minimize the distance. Given these $\{\nu_i\}_{i=1}^n$, it chooses for every $1 \le i \le n$

$$\lambda_i = \underset{\mu}{\operatorname{argmin}} \left\{ (1-\alpha)\frac{1}{d_P}\|x_i - f(\mu)\|^2 + \alpha \frac{6}{(n^2-1)}\left(d_i(\mu)^2 - d_i(\nu_i)^2\right) \right\}$$

that is, the λ_i that minimizes the weighted sum of the distance and the difference in contribution to the correlation term, given that all the other λ's remain the same. Notice that I divided the first term by d_P and multiplied the second by n, so that both terms will be independent of d_P and n, but this does not harm the optimization as it can be introduced back via α.

As noted in[169], Hastie and Stuetzle's algorithm can be analyzed from a probabilistic point of view. A principal curve with cubic spline smoothing is the solution for maximizing the likelihood that the curve generated the given data, assuming Gaussian noise and a prior on the smoothness of the curve. My modification, therefore, can be viewed as adding another prior, according to which the probability of a curve with $\{\lambda_i\}_{i=1}^n$ is

$$\Pr(\lambda) = e^{-\frac{6 d_P \alpha \left(d_i(\lambda_i)^2 - d_i(\nu_i)^2\right)}{(1-\alpha)(n^2-1)}}$$

In addition, since some ranking is known, I set the initial guess to be piecewise linear, generated by first calculating the average of all the points with the same rank, and connecting the averages corresponding to consecutive ranks. If the rank of most points is unknown, this might be a poor guess and I therefore start with the first principal component, making sure that its direction does not impose a reverse order to that of the ranking.

Setting the weight α

In order to automatically learn a suitable value of α, I propose the following heuristic. Find the best curve for a set of values of $\alpha \in [0,1)$ (say, in steps of 0.05) and record for each one the sum of squared distances of the data points to the curve, $D(\alpha)$, as well as the Spearman correlation $\rho(\alpha)$ between $\{\lambda_i\}_{i=1}^n$ and the labels. Ideally, I would like

to minimize the distance and maximize the correlation. However, as that is unlikely to occur simultaneously, I need to assign these two demands some weights.

I can therefore weigh the minimization goal $g(\alpha,R)=D(\alpha)-R\rho(\alpha)$, so that the right choice of R will allow selecting the best α. One possibility is to set $R=\frac{\alpha}{1-\alpha}$, or any other similar function of α, but this might lead to over or under-estimation of α – for example, if the maximal correlation is low, it would choose R=0, so that the correlation will not affect the target function, but this will force to set α=0, even though a higher value may be more appropriate, as it will improve the correlation with almost no effect on the distance. Therefore, two terms need to be to weighed in a way that is independent of α, yet flexible enough to handle those cases when both terms can be minimized considerably, as well as cases when a low enough minimum for one of the terms cannot be found. To achieve this, I minimize the difference between the standardized terms, that is:

$$\alpha=\underset{0\le\alpha<1}{\operatorname{argmin}}\left\{\frac{D(\alpha)-\bar{D}}{\sqrt{Var(D)}}-\frac{\rho(\alpha)-\bar{\rho}}{\sqrt{Var(\rho)}}\right\}$$

The mean and the variance are taken over the different values achieved for the different α's.

In many cases one might wish to use the ranking information only to guide the curve learning process, and not to set the actual final projection. To achieve that Pathifier can simply perform an additional iteration with no ranking information ($\alpha=0$). One reason one would want to do this is that often the ranking information cannot be trusted, or may be irrelevant to the gene set in question. Another related reason is that since one already knows the guiding ranking, new information may be more interesting, even at the cost of losing consistency with the guiding ranking. A more practical reason might be that I'm interested in using the curve to score new samples, for which I don't have any ranking information (for example using patient survival as the guiding ranking, and then predicting survival for a new patient on the basis of its projection onto the curve). In this the case I would like to assess the curve's performance in the same framework in which it is meant to be applied; that is, with no available ranking information.

A simulated example for semi-supervised principal curves

Suppose we are measuring the location of a particle over time. Our location measurements are a little noisy, and our time measurements are very noisy, to the extent we can only call whether the measurement was done at time *t=0* or *t=1* (e.g. the particle moves very fast). We would like to deduce the trajectory from the data. Approximately linear trajectories are easy to deduce, but complex ones require using principal curves or similar methods. In our toy example we have chosen the trajectory $(r\sin\theta, r\cos\theta)$ with Gaussian noise with standard deviation 0.3, where $r \in [1,3]$ and $\theta \in [0, 2\pi]$.

As expected, an unsupervised principal curve fails to capture the correct trajectory (see Appendix B). A possible way to introduce the knowledge of time could be in the initial guess. Accurate enough time measurements indeed allow the principal curve to get stuck in the correct local minimum. However in our case where time measurement is vague, a good initial guess is not enough – stating from the line connecting the center of the *t=0* points and the *t=1* points doesn't help. The modified algorithm handles this well, whether starting from the first principal component or not (see Appendix B).

Implementation and Availability

The code is implemented in R and Fortran based on "princurve 1.1-10" by Andreas Weingessel (http://cran.r-project.org/web/packages/princurve/index.html), which is in turn based on the original S/Fortran code princurv by Hastie. The code is available upon request.

Discussion

Pathifier performs pathway-level analysis of an expression dataset of tumors, and determines for each sample a set of pathway deregulation scores. These PDS are calculated separately for each pathway using genes which are known to take part in its functioning. The approach can be extended to any other kind of data with known pathway assignments, and allows incorporation of an additional parameter that is likely to be correlated with pathway deregulation (chromosomal instability, mutations, etc.). This framework may serve as basis for future integration of different measurements (such as DNA copy number, protein abundance, localization, etc.). The

approach is data-based: for each pathway Pathifier constructs a principal curve, which captures the variation of the data. All samples are projected onto this curve, and for each sample the distance between this projection and that of the normal samples is measured along the curve. This distance represents the level of deregulation of the pathway. The method copes successfully with the biggest challenges of expression-based pathway analysis: (a) knowledge of biological pathways is partial (b) pathway deregulation is context specific, and (c) available data (e.g. expression) contain only part of the relevant information. By including genes that were labeled by different studies as belonging to a pathway, and using data from the very problem we wish to study, we are able to define a context-specific PDS. This is accomplished without relying on (incompletely known) underlying network connectivity and function, by deducing pathway deregulation from the data itself. I deal with the absence of relevant information (e.g. post-translational modifications, protein localization) by projecting the very complex (and unavailable) parameterization of the "biological state" onto expression space, where deregulation is defined by the deviation from the signature of normal samples, measured along a trajectory that reflects context specific deregulation.

The main conceptual aim of defining these scores is to incorporate a large body of prior biological knowledge (in the form of assignment of genes to pathways) to allow further analysis on a "higher" (pathway) level, instead of analyzing directly the expression levels of tens of thousands of genes, in a brute force, "ignorance based" manner.

We applied the method to glioblastoma and showed that the PDS successfully reflect deregulation of pathways, and constitute a compact and biologically relevant representation of the samples. The resulting representation of the tumors retains most of the essential information present in the original data. We stratified tumors into subtypes which are easy to interpret in terms of the biologically meaningful and relevant pathways. The resulting tumor groups were consistent with previously identified clinical classes of glioblastoma and colon cancer, and we identified also new sub-classes with important clinical consequences.

For glioblastoma I found a clinically relevant new sub-stratification of Neural and Proneural samples, separating them into poor and good survivors, as well as a robust new substratification of the Mesenchymal subtype. I have shown that important recurrent mutations in glioblastoma have a clear impact on the deregulation scores of

the relevant pathways. Some samples without the mutation exhibit similar deregulation profiles to the mutated ones, suggesting alternative equivalent deregulation mechanisms. I found 35 pathways whose deregulation score is significantly correlated with survival, where higher levels of deregulation match poor survival. Some of these pathways (such as MAP kinase) were previously known to be associated with survival in glioblastoma patients, while several others constitute new findings (such as PDGFRβ and WNT signaling), that may serve as new hypotheses for glioblastoma research.

7. Differential peripheral-blood gene expression in patients with acute myocardial infarction and severe obstructive coronary atherosclerosis

In collaboration with Dr. Doron Aronson and Sagi Nahum of the Rappaport Faculty of Medicine and Research Institute at the Technion. Manuscript #10 (in preparation)

Introduction

Coronary *atherosclerosis* (hardening of the arteries) is caused when fat, cholesterol, and other substances build up in the walls of arteries and form hard structures called *plaques*. *Acute myocardial infarction* (MI, commonly known as heart attack) occurs when localized myocardial ischaemia (arrest of blood flow to heart muscle) causes the development of a defined region of necrosis. MI is most often caused by rupture of an atherosclerotic lesion in a coronary artery. This causes the formation of a thrombus that plugs the artery, stopping the flow of blood to the region of the heart that it normally supplies[178] (Figure 7-1). Coronary plaques which are prone to rupture are typically small and nonobstructive, with a large lipid-rich core covered by a thin fibrous cap. Activated macrophages and T-lymphocytes localized at the site of plaque rupture are thought to release metalloproteases and cytokines which weaken the fibrous cap, rendering it liable to tear or erode due to the shear stress exerted by the bloodflow. Severe and prolonged ischaemia depriving the corresponding heart muscle cells of oxygen produces a region of necrosis spanning the entire thickness of the myocardial wall. Such a transmural infarct usually causes *ST segment elevation MI* (STEMI, diagnosed when ECG shows elevation in the interval between ventricular depolarisation and repolarization, known as the ST segment, characteristic of heart muscle damage). Less severe and protracted ischaemia may cause non-ST segment elevation MI, NSTEMI.

Although all cases of plaque rupture occur in patients who have coronary atherosclerosis, most patients with atherosclerosis do not develop plaque rupture. Mortality and morbidity due to atherosclerosis are mostly related to MI rather than to atherosclerosis plaque growth and luminal narrowing[179-181]. Rupture or superficial erosion of vulnerable coronary atheroma leading to acute events do not correlate with prior severity of coronary *stenosis* (i.e., narrowing of the blood vessels caused by the

atherosclerosis)[180, 182, 183]. Indeed, atherosclerosis is a heterogeneous disease and some patients may progress to severe diffuse coronary luminal narrowing without ever developing acute coronary events. These clinical observations are in line with the fact that myocardial infarction and coronary artery disease (CAD) seem to be determined by different sets of genetic variations[184]. In this study we have aimed to use genomic profiling to better understand the molecular mechanisms that differentiate the general population of patient with CAD (i.e. severe stenosis) and those with STEMI, and to come up with candidates for biomarkers for STEMI for future research.

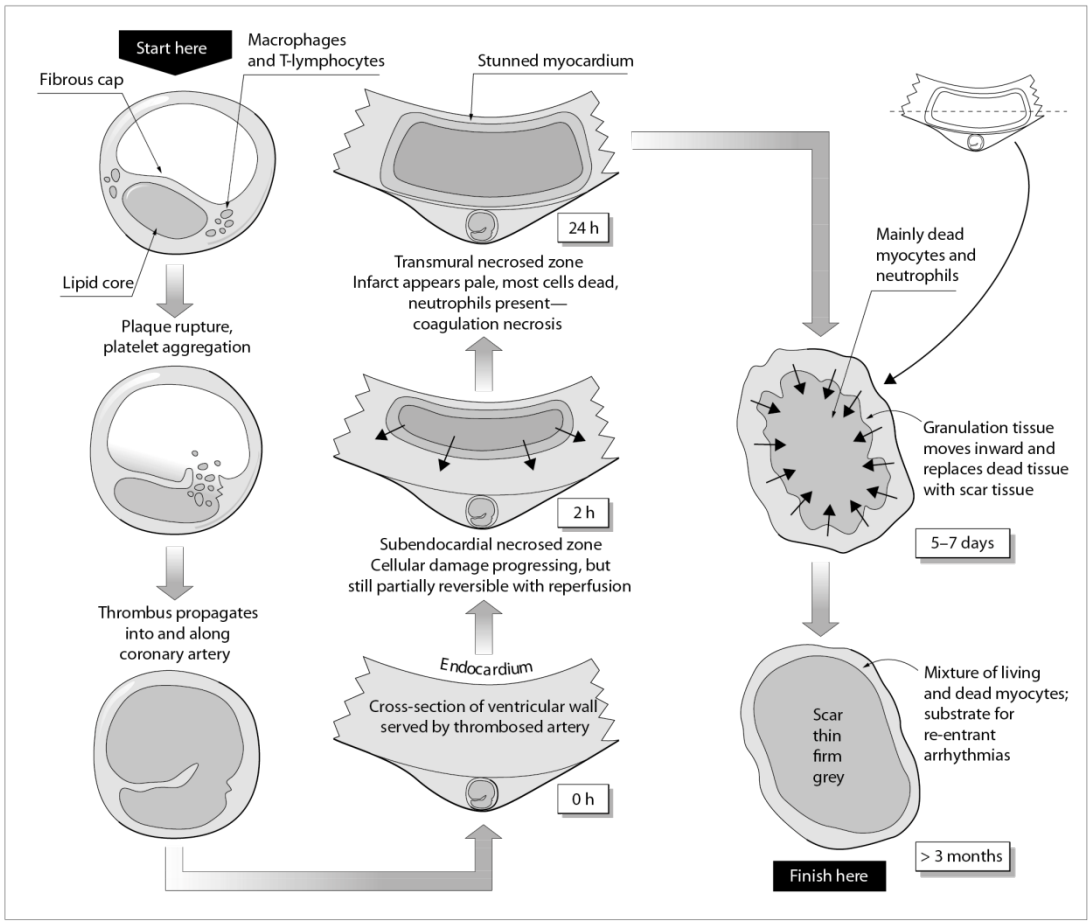


Figure 7-1 – **The development of acute myocardial infarction**, taken from[178]

The study of gene expression in cardiovascular disease involves the use of peripheral blood cells. This is because, unlike oncological tissue, the RNA of direct interest (i.e., vessel wall, fat tissue, and heart) is not readily available. Circulating leukocytes serve as a vigilant and comprehensive surveillance system that patrols the body for signs of infection and inflammation. Circulating blood cells are in contact with the diseased

endovascular lumen and as such, their gene expression profile seems to mirror events occurring in vascular tissue[185].

Recent studies have shown that patients with and without angiographically defined significant CAD may be distinguished by gene expression analysis of peripheral-blood cells[186-188]. However, it is unclear whether gene expression in peripheral blood can reflect overall disease burden or susceptibility for coronary plaque rupture and acute coronary events. The present study was designed to test the hypothesis that transcriptional variants in peripheral-blood of patients with coronary plaque rupture are distinct from those that are associated with the presence of coronary atherosclerosis. To this end, whole-genome expression profiling in circulating leukocytes was performed using coronary angiography as the primary discriminatory criterion for the various CAD phenotypes.

This work is the basis of a manuscript in preparation, appearing as #10 of the list presented in chapter 9.

Methods

Patients

Patients and control subjects were prospectively recruited from individuals that had undergone catheterization at the Rambam Health Care Campus Cardiac Catheterization Laboratory or underwent cardiac computed tomography. The investigational review committee on human research approved the study protocol and each patient signed an informed consent.

Patients were categorized into 3 groups: 1) MI: Heart attack patients who underwent primary percutaneous intervention for acute ST-elevation myocardial infarction with angiographic evidence of plaque rupture[189-191] in a single coronary vessel but with otherwise normal or near normal non-culprit vessels. 2) MV-CAD: Patients with severe multivessel CAD (but without plaque rapture), defined as ≥70% stenosis in at least 2 major epicardial coronary vessels by quantitative analysis[192] but without clinical history, electrocardiographic or echocardiographic evidence of myocardial infarction and without angiographic characteristics of unstable coronary plaques[189-191]. 3) Healthy subjects (no CAD, no MI) with CAD risk factors and angiographically normal coronary (NC) arteries or a cardiac computed tomography demonstrating

absence of coronary plaques with coronary artery calcium score of zero, and without any history of peripheral arterial or cerebrovascular disease.

Exclusion criteria included 1) renal insufficiency (creatinine > 2.0 mg/dL), 2) present or past history of malignancy, 3) significant valvular heart disease, 4) heart failure, 5) cigarette smoking >2 packs per day, 6) anemia (hemoglobin <12.0 g/dL for females or <13.0 g/dL for males), 7) systemic inflammatory disease and 8) use of any immunosuppressive agents.

Gensini's score was used to assess the extent of CAD[193]. We modified the original scoring system by adding a specific score for acute total occlusion, as in the case of acute MI. As acute coronary occlusion usually occurs in a previous angiographically non-critical lesion[182], we scored acute total occlusion as a non-significant lesion (from 0 to 5 score) instead of true chronic total occlusion (from 32 to 172 score) as previously described[194]. As part of the protocol, all blood samples were obtained 4-6 months after the infarction or revascularization procedure, a period adequate for any residual effects of acute tissue infarction/ischemia on systemic inflammation to have disappeared[195, 196].

Cell collection

Peripheral blood mononuclear cells (PBMCs) were obtained from heparinized blood by Isopaque-Ficoll (Lymphoprep, Nycomed, Oslo, Norway). After gradient centrifugation at room temperature (800g for 20 min), the PBMCs were washed twice with phosphate-buffered saline (PBS) and RNA was extracted immediately.

RNA preparation

Total cellular RNA was extracted from PBMC using the Qiagen RNAeasy Kit (QIAGEN, Chatsworth, CA) according to manufacturer's protocol. The RNA samples were treated with DNAse (RNase free DNase set; QIAGEN) to prevent genomic DNA amplification. RNA quantity and quality were determined using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and by Bio-Rad Experion Automated Electrophoresis Station (Bio-Rad, Hercules, California, USA).

Microarray processing

For screening differential gene expression between patients and control groups, we used Illumina Human-6 Expression BeadChip system (Illumina Inc., San Diego, CA)[197]. The microarray gene chip Human WG-6 version 3 contains 47,289 unique 50-mer oligonucleotides in total, with hybridization to each probe assessed at ~30 different beads on average. For gene expression analysis, 200ng of total RNAs were reverse transcribed and first- and second-strand complementary DNA (cDNA) were synthesized. After cDNA purification, biotin-labeled complementary RNA (cRNA) was synthesized using Illumina TotalPrep RNA Amplification Kit (Applied Biosystems/Ambion, Austin, USA). A total of 1.5 μg of biotin-labeled cRNA was hybridized on the chip at 55°C for 18h. The hybridized BeadChips were washed and labeled with streptavidin-Cy3.

The samples were scanned on the Illumina BeadArray 500GX Reader using Illumina BeadScan image data acquisition software (version 2.3.0.13). Quality control and quantile normalization of the microarray data was done by BeadStudio 3.0 software (Illumina). To ensure reproducibility of the data and to allow for variation calculation, we included on each BeadChip a technical replicate consisting of a single recurrent RNA sample, as well as a biological replicate, namely a sample that was separately collected from the patient being assessed.

Validation by quantitative real-time polymerase chain reaction

Primers were designed by using the Primer Express 2.0 software and verified by using a BLAST search. ~200ng RNA were used to synthesize cDNA at 47°C using the Verso TM cDNA kit (ABgene, UK). Taqman probes (Egr-1, JUNB, SERPINA1, SNHG5, OSM) were provided by Applied Biosystems. Primer sets for use with SYBR Green (IL-1β, FOLR3, RSP26, Egr-2, Cox-2) were designed using Primer Express software (Applied Biosystems). Amplification reactions were performed in duplicates from 20ng cDNA using the Applied Biosystems 7300 Real Time PCR System (Warrington, Cheshire, UK) with the following cycling parameters: 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min. SYBR green amplification reactions were performed using the Mx4000 Multiplex Quantitative PCR System (Stratagene, La Jolla, CA, USA) under the following conditions: 50°C for 2 min, 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min.

ACTB and GAPDH genes were used as normalization genes. Relative gene expression was assessed by the $2^{-\Delta\Delta CT}$ method[198].

Microarray Data processing

Data was stabilized based on the approach by Lin et al.[199]. To retain information on the variance, typically lost in the Lin et al. process, the variance stabilizing transformation was calculated on the summarized data, but applied on the raw data (after omitting outliers at 4 median absolute deviations[MADs] away from the median). This allowed the calculation of the variance after the transformation.

The summarized data were normalized with quantile normalization to minimize technical effects. The normalization is first done between each two strips of the same sample, since we assume they should be identical up to the strip effect, and then once again on all samples from the same batch. Distance Weighted Discrimination[200] was used to correct for batch effects between the three batches of the experiment. Absent/Present call was based on the detection p-value calculated from the intensities following Illumina's protocol[201], and taking the minimal p-value of the two strips. Absent probes were filtered out by requiring each probe to pass a detection FDR of 10% for at least 5 samples, leaving 14,734 probes. Of those only the 10,000 most varying probes were used for the analysis to lower the noise level.

Statistical analysis

Modified one-way ANOVA followed by Tukey's post-hoc test was applied to reveal genes that demarcate the three sample groups. The modification did not allow the variance taken for each probe, typically calculated across each sample type, to be less than the (technical) variance estimated for that probe (across the repeats of the probe in each chip). Probes with FDR<0.25 and fold change ≥1.3 were recorded. Following the ANOVA procedure, Tukey's post-hoc tests were performed to detect differences in group classification membership. A Venn diagram was created, with genes that showed a minimum fold change of 1.3-fold in each comparison. Genes and samples (within each sample set) were sorted by SPIN[202]. GO annotation were calculated using DAVID[203, 204].

Results

Microarray Analysis

Comparative analysis revealed 298 genes that were differentially expressed (fold change >1.3, FDR<0.25). Of these, 166 genes showed significant difference between MV-CAD and NC, of which 136 genes were expressed at higher levels in the MV-CAD patients and 30 were down-regulated. There were 171 differentially expressed genes between MV-CAD and MI patients (118 higher in MV-CAD, 53 in MI). One hundred genes were differentially expressed genes between MI patients and subjects with NC (32 up-regulated, 68 down regulated). Hierarchical clustering and visualization of the expression levels of the 298 differentially expressed genes is shown in Figure 7-1 below.

Three main clusters can be recognized, corresponding to the 3 study groups: cluster I, of 72 genes highly expressed in NC patients and with low expression in the MI patients; cluster II, of 71 genes with a higher expression in MI compared with NC and MV-CAD; and cluster III, of 155 genes that show mainly high expression in MV-CAD patients as compared with MI and NC.

Ninety two genes were significantly differentially expressed between the MV-CAD vs. NC *and* MV vs. MI comparisons. The great majority of these gens (n = 85, 92%) showed higher expression levels in MV-CAD and MI patients than in NC subjects, and belong to cluster III which contained transcripts with an established role in the pathogenesis of atherosclerosis (e.g., IL-1β, EGR1, PTGS2, OSM, NAMPT, TLR6, TLR4, SOD2). Because these genes separated patients with stable MV-CAD (severe atherosclerosis) from *both* patients with MI (mild atherosclerosis) and from study participants without atherosclerosis, this group of genes may regulate processes related to the progression of atherosclerosis (rather than plaque rupture).

Sixteen genes were significantly differentially expressed between MV-CAD vs. MI and MI vs. NC, 10 of which higher on MI (belong to cluster II) and 6 lower (belong to cluster I). Similarly, 26 genes were significantly different between MV-CAD vs. NC and MI vs. NC, 14 of which higher in NC (cluster I) and 12 lower (cluster II or III).

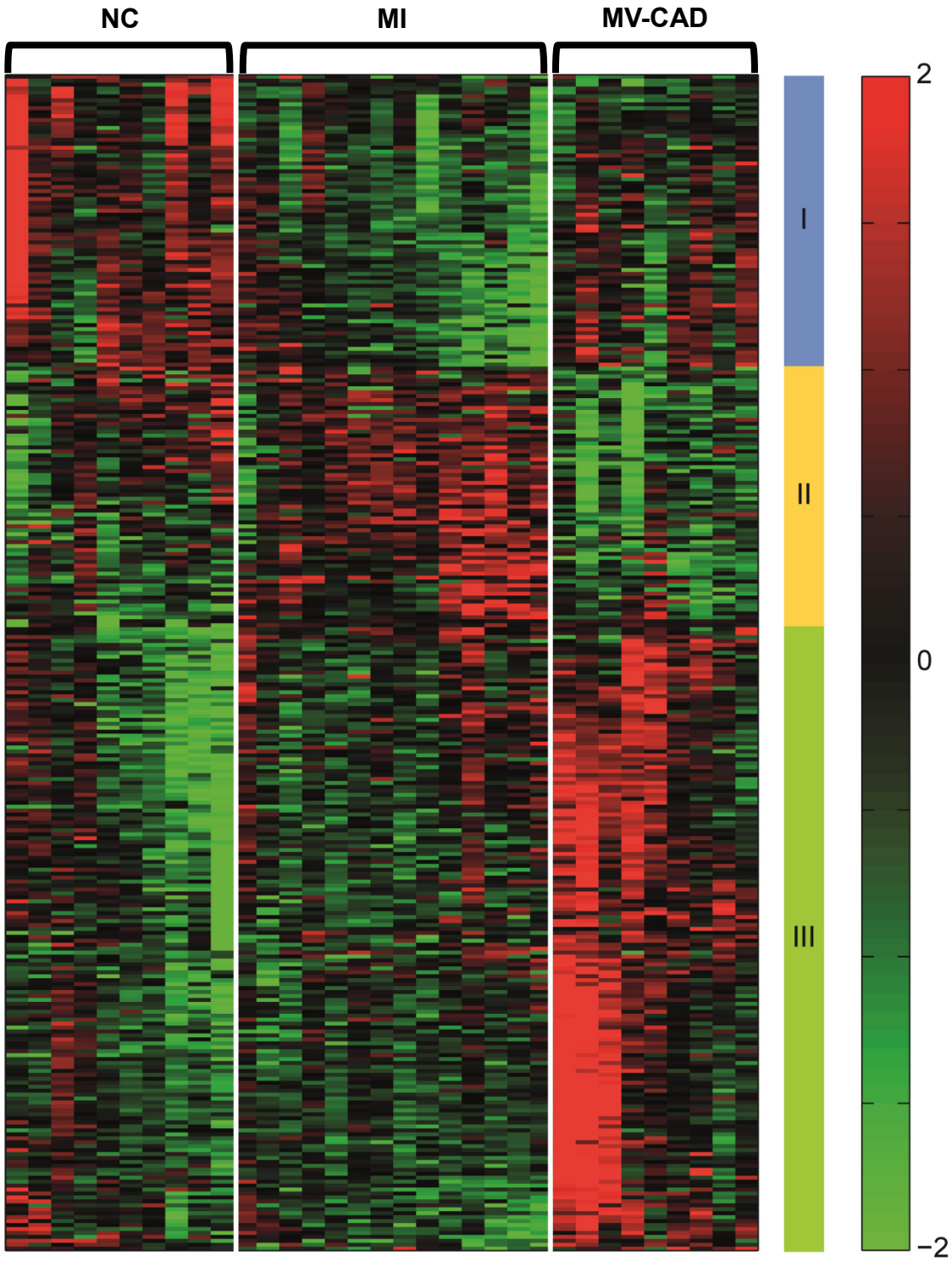


Figure 7-1 - **Differentially expressed genes between the three patient groups** (genes altered >1.3-fold; *FDR*<25%, one-way ANOVA). Columns represent individual patients, and rows individual genes. Gene expression is represented as a color, with green representing below-average expression level and red representing above-average expression level. The bar on the left denotes the three clusters. Genes and samples (within each sample set) were sorted by SPIN[202].

Validation of Microarray With Quantitative PCR

To confirm the microarray findings, we carried out quantitative PCR (qPCR) analysis of gene expression using the original patient population and 10 additional participants (5 patients with MV CAD and 5 with MI). We selected 10 genes of potential biological interest for validation of the microarray findings by quantitative PCR assay with either TaqMan assays or SYBR Green.

As shown in Figure 7-2, qPCR confirmed differential abundance of the top microarray-determined expression markers between patient classes. The qPCR gene expression results carried out on the same set of samples that were analyzed by the microarray approach were highly correlated with those from the microarray data. The correlation coefficients for microarray and qPCR for the 10 selected genes were 0.74 to 0.99.

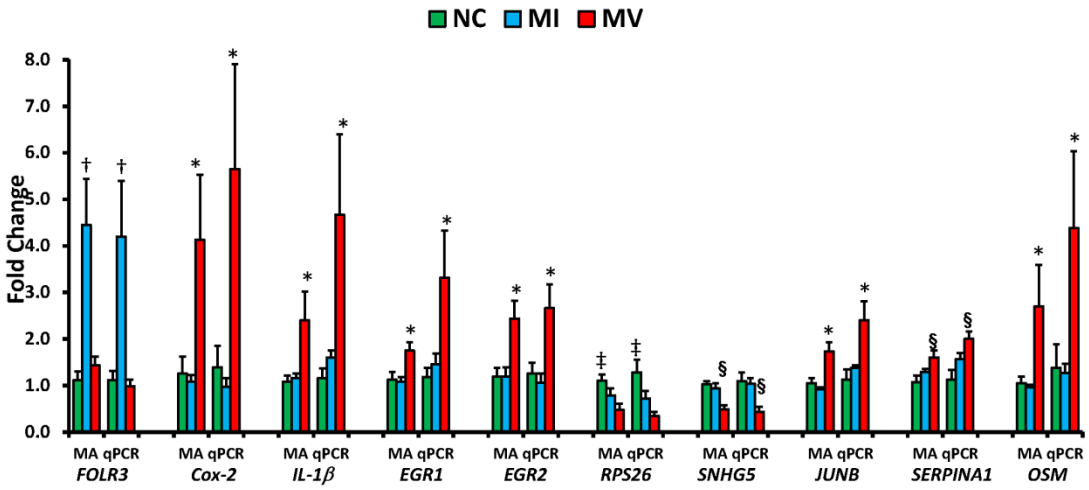


Figure 7-2 - **Confirmation of microarray (MA) result by quantitative PCR (qPCR) in the 3 study groups**. The bars show the microarray (MA) and qPCR fold change of the 10 selected genes (Cox-2, Egr-1, IL-1β, Egr-2, FOLR3, RSP26, JUNB, SERPINA1, SNHG5, OSM). To standardize the amount of input RNA, ACTB and GAPDH genes were selected as normalization genes. Relative gene expression was assessed by the $2^{-\Delta\Delta CT}$ method. **P* < 0.05 for MV compared with MI and with NC; †*P* < 0.05 for MI compared with MV and NC; ‡*P* < 0.01 NC vs. MV; §*P* < 0.01 MV vs. MI and NC.

Pathway Analysis

We have analyzed the pathways that are enriched in each cluster, complete results (FDR<1%) are in Table 7-1 below. Cluster I is mainly enriched in Histones, as shown in an enrichment of chromatin and nucleosome organization pathways. Cluster II is enriched in RNA binding proteins. Cluster III is enriched in a large number of pathways, many of which are related to immune response. As this cluster is of genes

highly expressed in MV-CAD patients, and the level of athrothrombosis is also expected to be higher on MV-CAD, and the genes are enriched in immune and inflammatory response, we suggest that many of these genes correlate with the level of athrothrombosis. To further investigate this, we have compared the patient Gensini score with median of the 17 probes of cluster III that belong to the inflammatory response pathway (GO:0006954). The correlation is indeed significant ($R=0.47$, $p<0.006$), see Figure 7-3.

Cluster I		Cluster II		Cluster III	
Term	p-Value	Term	p-Value	Term	p-Value
GO:0032993 protein-DNA complex	2.58E-10	GO:0003723 RNA binding	2.85E-04	GO:0006955 immune response	4.59E-10
GO:0000786 nucleosome	9.87E-10			GO:0006952 defense response	2.18E-08
GO:0006334 nucleosome assembly	2.88E-09			GO:0006954 inflammatory response	4.41E-07
GO:0031497 chromatin assembly	3.75E-09			GO:0031328 positive regulation of cellular biosynthetic process	3.62E-06
GO:0065004 protein-DNA complex assembly	5.26E-09			GO:0009891 positive regulation of biosynthetic process	4.37E-06
GO:0034728 nucleosome organization	6.18E-09			GO:0031224 intrinsic to membrane	4.64E-06
GO:0006323 DNA packaging	3.35E-08			GO:0009611 response to wounding	6.85E-06
GO:0006333 chromatin assembly or disassembly	6.07E-08			GO:0010557 positive regulation of macromolecule biosynthetic process	7.36E-06
GO:0034622 cellular macromolecular complex assembly	2.51E-07			GO:0005886 plasma membrane	1.49E-05
GO:0034621 cellular macromolecular complex subunit organization	6.76E-07			GO:0010033 response to organic substance	2.71E-05
GO:0000785 chromatin	3.34E-06			GO:0002237 response to molecule of bacterial origin	8.25E-05
GO:0065003 macromolecular complex assembly	1.59E-05			GO:0016021 integral to membrane	9.81E-05
GO:0043933 macromolecular complex subunit organization	2.81E-05			GO:0009617 response to bacterium	1.50E-04
GO:0044427 chromosomal part	2.91E-05			GO:0010604 positive regulation of macromolecule metabolic process	2.48E-04
GO:0051276 chromosome organization	6.64E-05			GO:0051173 positive regulation of nitrogen compound metabolic process	2.97E-04
GO:0006325 chromatin organization	9.61E-05			GO:0042035 regulation of cytokine biosynthetic process	3.84E-04
GO:0005694 chromosome	1.01E-04			GO:0050727 regulation of inflammatory response	4.09E-04
GO:0003924 GTPase activity	3.31E-04			GO:0032496 response to lipopolysaccharide	4.62E-04
GO:0005525 GTP binding	6.07E-04			GO:0051384 response to glucocorticoid stimulus	4.90E-04
GO:0032561 guanyl ribonucleotide binding	7.00E-04				
GO:0019001 guanyl nucleotide binding	7.00E-04				

Table 7-1 – **GO annotations of the clusters** of Figure 7-1 that passed 1% FDR.

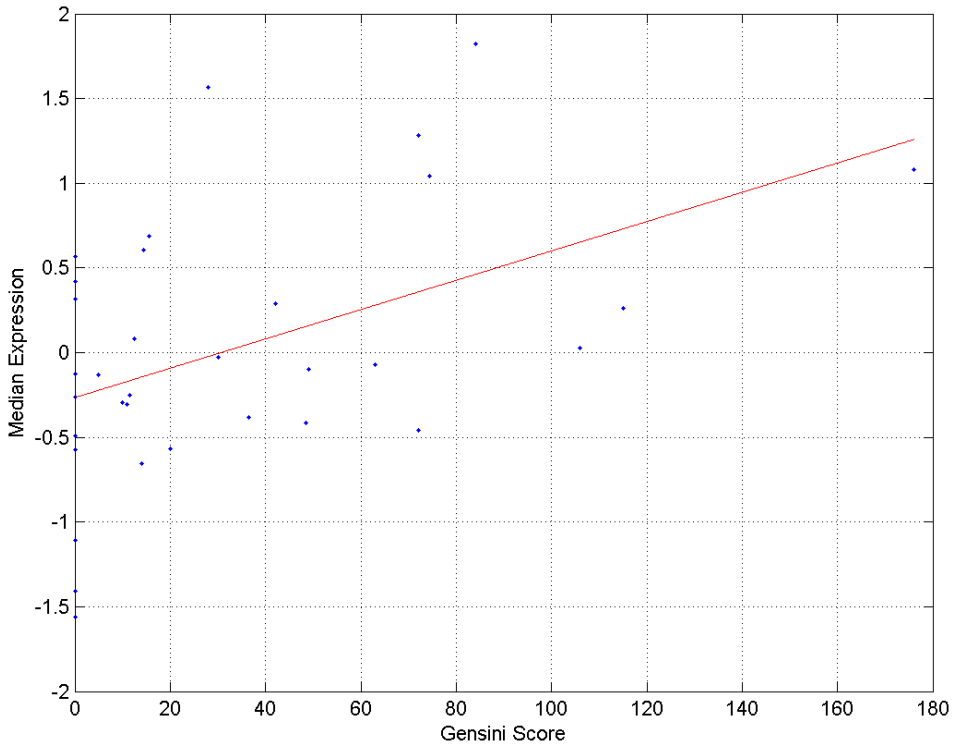


Figure 7-3 - **Median expression of inflammatory genes of cluster III correlates with level of athrothrombosis**. The 17 probes used match the following 15 genes: AIF1, CD55, CEBPB, CLEC7A, FPR2, IL15, IL1B, IL1RN, LY96, RXRA, SERPINA1, TLR4, TLR6, TNFAIP6, TNFRSF1A.

Discussion

In the present study, we set out to uncover transcriptional determinants of human coronary atherosclerosis and coronary plaque rupture. The use 3 distinct CAD phenotypes reduced heterogeneity in coronary atherosclerosis burden within patients and allowed discrimination of transcription patterns characteristic of the severity of atherosclerosis from those associated with acute coronary events. The principal findings of the present study are: 1) Severe stable MV-CAD is associated with the activation of numerous genes and inflammatory pathways in PBMCs which partly overlap with genes known to be expressed within the diseased human vascular wall, with an overall pattern consistent of systemic inflammation; 2) Our data demonstrates differential gene expression in PBMCs of patients with a history of coronary plaque rupture and patients with stable MV-CAD, with immune activation predominating in the latter group; 3) Both transcriptional information of individual genes and pathway

analysis indicated that the severity of atherosclerosis correlates with immune activation. MI patients, who by study design had lower coronary plaque burden than MC–CAD patients, demonstrated an intermediate inflammatory activation, with lower activation compared to MV-CAD and higher activation as compared with subjects without coronary atherosclerosis.

Several large studies have shown a poor association between burden of atherosclerosis and soluble biomarkers of inflammation such as C-reactive protein[205, 206]. In contrast, the present study demonstrates that transcriptomic information robustly separates patients with clinically significant obstructive CAD from subjects without CAD. These data suggests that that transcriptomic information is more sensitive than soluble markers of inflammation in identifying subjects with clinically significant CAD, and raises the possibility that gene expression in PBMCs may be an excellent marker of ongoing systemic inflammation.

Many of the genes identified as differentially expressed were directly relevant to the pathogenesis of the atherosclerotic process. Importantly, transcripts of molecules known to be present in the human coronary plaque and to play an important role in atherosclerosis progression were found to be overexpressed in peripheral blood of patients with MV–CAD. For example, PBMCs of patients with MV-CAD exhibited increased expression of early growth response genes (Egr-1 and Egr-2) encoding zinc-finger transcription factors that modulate a cluster of stress-responsive genes including platelet derived growth factor and transforming growth factor-β. Egr-1 has been shown to be differentially regulated in several cell types within atherosclerotic lesions and is thought to be a factor required in atherogenesis[207, 208]. Egr-1 is also found within inflammatory cells of vascular lesions, such as $CD68^+$ macrophages of aortic atherosclerotic lesions[209]. Similarly, IL-1β is a prototypic inflammatory cytokine, that has been shown to localize in foam cells, smooth muscle cells, and endothelium in atherosclerotic plaques[210, 211]. Oncostatin M (OSM), a macrophage and T-lymphocyte specific inflammatory cytokine and members of the gp130 cytokine family, is also expressed in macrophages and smooth muscle cells within human atherosclerotic lesions[212-214]. OSM seems to be involved in progression, matrix remodeling, and calcification of atherosclerotic lesions. COX-2 is induced in atherosclerotic plaques but not the normal blood vessels, and has been localized in proliferating vascular smooth muscle cell and particularly in inflammatory cells[215, 216]. JunB (AP-1) is regulated by several growth factors, cytokines, and by various agents

that induce oxidative stress and is involved in PDGF-induced vascular smooth muscle cell migration and proliferation[217] and in the regulation of VCAM-1[218] and tissue factor[219] expression in endothelial cells.

Overall, differentially expressed genes included predominantly members of innate immune signaling pathways (TLR4, TLR6, IL-1β, IL1RN, IL-15, Visfatin, S100), but also genes related to adaptive immunity (IL-15, LAT). Interestingly, patients with MV-CAD also demonstrated increased expression of apoptosis-promoting transcripts including TRAIL, KILLER/DR5, TNFRSF8 and IFIT2[220, 221]. Increased expression of TRAIL in PBMCs is consistent with previous reports of TRAIL expression in human atherosclerotic plaques[222].

Thus, gene expression pattern in PBMCs not only reflected a general immune activation but also demonstrated changes in expression of specific molecules known to be expressed in the diseased vessel wall, with specific role in the pathobiology of atherosclerosis. Taken together, these findings support the concept that gene expression patterns in PBMCs may mirror biological processes within the diseased vascular wall[185] with implications for the noninvasive detection of CAD[188].

Pathway analysis: The most notable finding of the pathway analysis was an enrichment of immune and inflammatory response in cluster III. Further examination revealed that indeed some of the genes in this cluster significantly correlate with the level of athrothrombosis (Figure 7-3). These results support the hypothesis that several inflammatory pathways are mainly linked to the progression of atherosclerosis, without a clear impact on the occurrence of coronary events.

The fact that many patients with severe and extensive atherosclerosis remain stable for years without developing acute coronary syndromes, while others develop acute events as the first manifestation of CAD despite less severe coronary atherosclerosis remains poorly understood[223]. A notable finding of the present study was that activation of multiple inflammatory pathways was more pronounced in patients with MV–CAD as compared with MI patients in the convalescent phase. The data are consistent with histopathologic studies showing that high-grade inflammatory infiltration is also present in stable plaques of patients with ACS irrespective of infarct-related segments and non-infarct-related segments[224].

The finding of higher immune activation in MV-CAD as compared with MI patients does not rule out the possibility of a transient elevation of inflammatory activity triggered just prior to an acute plaque rupture event[225]. As blood samples were

obtained 4-6 months after the event, the lower inflammatory activation in MI patients may be explained by lesion stabilization after rupture and symptomatic acute MI, with overall decrease in inflammatory activity with time elapsed between the event and blood sampling, with the remaining inflammatory activity related to the "background" atherosclerosis. However, our findings also indicate that in stable patients, inflammation in PBMCs is related to the extent of plaque burden. Importantly, these results also demonstrate that patients with MV–CAD may have pronounced inflammatory activation for prolonged periods of time without ever experiencing an acute coronary event. Thus, because robust systemic inflammation is not inconsistent with long-term stability of severe obstructive CAD, inflammation may also play a role in plaque stabilization. Indeed, it has been suggested that the inflammatory stimuli that trigger plaque disruption might differ from the inflammatory stimuli that promote plaque growth[223].

Conclusion: The results of the present study demonstrate that PBMCs expression profile of patients with CAD is characterized by differential expression of specific molecules and activity of molecular networks that have functional significance for atherogenesis. Inflammatory gene expression in PBMCs mainly reflects the severity of atheromatous burden. Robust systemic inflammation is not inconsistent with long-term stability of severe obstructive CAD, suggesting that some inflammatory pathways may play a role in plaque stabilization.

8. Two phases of mitogenic signaling unveil roles for p53 and EGR1 in elimination of inconsistent growth signals

In collaboration with Prof. Yosef Yarden and Dr. Yaara Zwang of Weizmann Institute. Publication #8.

In this work we have integrated comprehensive proteomic and transcriptomic analyses of cells given two EGF pulses. These two pulses, when appropriately spaced commit the cells to cell-cycle progression. Through the analysis we identified two gating mechanisms, which ensure that cells ignore fortuitous growth factors and undergo proliferation only in response to consistent mitogenic signals: The first pulse induces essential metabolic enzymes and activates p53-dependent restraining processes; the second pulse eliminates, via the PI3K/AKT pathway, the suppressive action of p53, as well as sets an ERK-EGR1 threshold mechanism.

The details have been published in Molecular Cell (publication 8 of the list presented in chapter 9).

9. List of Ph.D. publications

1. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel S, Lander ES, Meyerson M, Beroukhim R, Getz G: Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. **Genome Research 2012**, Accepted (10.1101/gr.141382.112).
2. Drier Y, Sheffer M, Domany E: *Non-linear estimation of pathway deregulation provides biologically relevant and compact representation of tumors*. Submitted.
3. Williams LJS, Tabbaa DG, Li N, Berlin AM, Shea TP, MacCallum I, Lawrence MS, Drier Y, Getz G, Young SK, Jaffe DB, Nusbaum C and Gnirke A, *Paired-end sequencing of Fosmid libraries by Illumina*. **Genome Research 2012**, 22(11):2241-9.
4. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, Ren X, Cibulskis K, Sivachenko AY, Wagle N, Sucker A, Sougnez C, Onofrio R, Ambrogio L, Auclair D, Fennell T, Carter SL, Drier Y, Stojanov P, Singer MA, Voet D, Jing R, Saksena G, Barretina J, Ramos AH, Pugh TJ, Stransky N, Parkin M, Winckler W, Mahan S, Ardlie K, Baldwin J, Wargo J, Schadendorf D, Meyerson M, Gabriel SB, Golub TR, Wagner SN, Lander ES, Getz G, Chin L, Garraway LA: *Melanoma genome sequencing reveals frequent PREX2 mutations*. **Nature 2012**, 485(7399):502-6.
5. Berger MF*, Lawrence MS*, Demichelis F*, Drier Y*, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA: *The genomic complexity of primary human prostate cancer*. **Nature 2011**, 470(7333):214-220.

   * Equal Authorship
6. Drier Y, Domany E: Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? **PLoS One 2011**, 6(3):e17795.

7. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, Jing R, Parkin M, Pugh T, Verhaak RG, Stransky N, Boutin AT, Barretina J, Solit DB, Vakiani E, Shao W, Mishina Y, Warmuth M, Jimenez J, Chiang DY, Signoretti S, Kaelin WG, Spardy N, Hahn WC, Hoshida Y, Ogino S, Depinho RA, Chin L, Garraway LA, Fuchs CS, Baselga J, Tabernero J, Gabriel S, Lander ES, Getz G, Meyerson M: *Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion*. **Nat Genet. 2011** 43(10):964-8.

8. Zwang Y, Sas-Chen A, Drier Y, Shay T, Avraham R, Lauriola M, Shema E, Lidor-Nili E, Jacob-Hirsch J, Amariglio N, Lu Y, Mills GB, Rechavi G, Oren M, Domany E, Yarden Y: *Two phases of mitogenic signaling unveil roles for p53 and EGR1 in elimination of inconsistent growth signals*. **Mol Cell. 2011** 42(4):524-35.

9. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR: *Initial genome sequencing and analysis of multiple myeloma*. **Nature 2011**, 471(7339):467-472.

10. Nahum S*, Drier Y*, Shmoish M, Avidan N, Sarig O, Domany E, Sprecher E, Aronson D: *Patients with acute myocardial infarction and severe obstructive coronary atherosclerosis display distinct peripheral blood gene expression profiles.* In preperation.

    * Equal Authorship

10.References

1. Hoeijmakers, J.H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-374 (2001).
2. DePinho, R.A. & Polyak, K. Cancer chromosomes in crisis. *Nat Genet* **36**, 932-934 (2004).
3. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729 (2008).
4. Stephens, P.J. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010 (2009).
5. Campbell, P.J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-1113 (2010).
6. Totoki, Y. et al. High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* **43**, 464-469 (2011).
7. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).
8. Chapman, M.A. et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
9. Bass, A.J. et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-968 (2011).
10. Wang, L. et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365**, 2497-2506 (2011).
11. Stransky, N. et al. The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science* (2011).
12. Berger, M.F. et al. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506 (2012).
13. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409 (2012).
14. Zhu, C. et al. Unrepaired DNA breaks in p53-deficient cells lead to oncogenic gene amplification subsequent to translocations. *Cell* **109**, 811-821 (2002).
15. Mills, R.E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).
16. Chiang, C. et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44**, 390-397, S391 (2012).

17. Kaplan, K.B. et al. A role for the Adenomatous Polyposis Coli protein in chromosome segregation. *Nat Cell Biol* **3**, 429-432 (2001).
18. Guerrero, A.A. et al. Centromere-localized breaks indicate the generation of DNA damage by the mitotic spindle. *Proc Natl Acad Sci U S A* **107**, 4159-4164 (2010).
19. Fearon, E.R. Molecular genetics of colorectal cancer. *Annu Rev Pathol* **6**, 479-507 (2011).
20. TCGA Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012).
21. De, S. & Babu, M.M. A time-invariant principle of genome evolution. *Proc Natl Acad Sci U S A* **107**, 13004-13009 (2010).
22. Kino, K. & Sugiyama, H. UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutat Res* **571**, 33-42 (2005).
23. Kino, K. & Sugiyama, H. Possible cause of G-C-->C-G transversion mutation by guanine oxidation product, imidazolone. *Chem Biol* **8**, 369-378 (2001).
24. Jansen, J.G. et al. Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. *J Exp Med* **203**, 319-323 (2006).
25. Ross, A.L. & Sale, J.E. The catalytic activity of REV1 is employed during immunoglobulin gene diversification in DT40. *Mol Immunol* **43**, 1587-1594 (2006).
26. Stephens, P. et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37**, 590-592 (2005).
27. Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
28. Rubin, A.F. & Green, P. Mutation patterns in cancer genomes. *Proc Natl Acad Sci U S A* **106**, 21766-21770 (2009).
29. Beale, R.C. et al. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J Mol Biol* **337**, 585-596 (2004).
30. Bishop, K.N. et al. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Curr Biol* **14**, 1392-1396 (2004).
31. Nik-Zainal, S. et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979-993 (2012).
32. Volik, S. et al. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A* **100**, 7696-7701 (2003).

33. Raphael, B.J., Volik, S., Collins, C. & Pevzner, P.A. Reconstructing tumor genome architectures. *Bioinformatics* **19 Suppl 2**, ii162-171 (2003).
34. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-732 (2005).
35. Bashir, A., Volik, S., Collins, C., Bafna, V. & Raphael, B.J. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* **4**, e1000051 (2008).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
37. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
38. Robinson, J.T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
39. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
40. Reich, M. et al. GenePattern 2.0. *Nat Genet* **38**, 500-501 (2006).
41. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).
42. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8**, 652-654 (2011).
43. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
44. Scholz, F.W. & Stephens, M.A. K-Sample Anderson-Darling Tests. *J Am Stat Assoc* **82**, 918-924 (1987).
45. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
46. Ryba, T. et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**, 761-770 (2010).
47. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
48. Chen, L. et al. Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma. *Blood* **115**, 61-70.

49. Johansson, P., Pavey, S. & Hayward, N. Confirmation of a BRAF mutation-associated gene expression signature in melanoma. *Pigment Cell Res* **20**, 216-221 (2007).
50. Pyeon, D. et al. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res* **67**, 4605-4619 (2007).
51. Kohlmann, A. et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol* **142**, 802-807 (2008).
52. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
53. Meister, P., Taddei, A. & Gasser, S.M. In and out of the replication factory. *Cell* **125**, 1233-1235 (2006).
54. Tillier, E.R. & Collins, R.A. Genome rearrangement by replication-directed translocation. *Nat Genet* **26**, 195-197 (2000).
55. Eisen, J.A., Heidelberg, J.F., White, O. & Salzberg, S.L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1**, RESEARCH0011 (2000).
56. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol* (2011).
57. Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**, 104-115 (2007).
58. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).
59. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
60. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-1065 (2011).
61. Meaburn, K.J., Misteli, T. & Soutoglou, E. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* **17**, 80-90 (2007).
62. Mani, R.S. & Chinnaiyan, A.M. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat Rev Genet* **11**, 819-829 (2010).

63. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L.A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol* (2011).
64. Klein, I.A. et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**, 95-106 (2011).
65. Farkash-Amar, S. & Simon, I. Genome-wide analysis of the replication program in mammals. *Chromosome Res* **18**, 115-125 (2010).
66. Yu, J. et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443-454 (2010).
67. Lin, B. et al. Integrated expression profiling and ChIP-seq analyses of the growth inhibition response program of the androgen receptor. *PLoS One* **4**, e6589 (2009).
68. Carroll, J.S. et al. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297 (2006).
69. Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).
70. Pleasance, E.D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
71. Lee, W. et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
72. Waterman, M.L. Lymphoid enhancer factor/T cell factor expression in colorectal cancer. *Cancer Metastasis Rev* **23**, 41-52 (2004).
73. Korinek, V. et al. Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. *Science* **275**, 1784-1787 (1997).
74. Kriegl, L. et al. LEF-1 and TCF4 expression correlate inversely with survival in colorectal cancer. *J Transl Med* **8**, 123 (2010).
75. Folsom, A.R. et al. Variation in TCF7L2 and increased risk of colon cancer: the Atherosclerosis Risk in Communities (ARIC) Study. *Diabetes Care* **31**, 905-909 (2008).
76. Hazra, A., Fuchs, C.S., Chan, A.T., Giovannucci, E.L. & Hunter, D.J. Association of the TCF7L2 polymorphism with colorectal cancer and adenoma risk. *Cancer Causes Control* **19**, 975-980 (2008).
77. Tuupanen, S. et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885-890 (2009).

78. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-884 (2009).
79. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274 (2006).
80. Wood, L.D. et al. The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113 (2007).
81. Parsons, D.W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812 (2008).
82. TCGA Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
83. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).
84. Tomlins, S.A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599 (2007).
85. Carver, B.S. et al. Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat Genet* **41**, 619-624 (2009).
86. King, J.C. et al. Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation in prostate oncogenesis. *Nat Genet* **41**, 524-526 (2009).
87. Han, B. et al. Fluorescence in situ hybridization study shows association of PTEN deletion with ERG rearrangement during prostate cancer progression. *Mod Pathol* **22**, 1083-1093 (2009).
88. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171-178 (2005).
89. Koscielny, S. Critical review of microarray-based prognostic tests and trials in breast cancer. *Curr Opin Obstet Gynecol* **20**, 47-50 (2008).
90. Kim, S.Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics* **10**, 147 (2009).
91. van't Veer, L.J. & Bernards, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* **452**, 564-570 (2008).
92. Drier, Y. & Domany, E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS One* **6**, e17795 (2011).

93. Bild, A.H. et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353-357 (2006).
94. Thomas, D.C. et al. Approaches to complex pathways in molecular epidemiology: summary of a special conference of the American Association for Cancer Research. *Cancer Res* **68**, 10028-10030 (2008).
95. Chin, L., Hahn, W.C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes Dev* **25**, 534-555 (2011).
96. Lewitter, F., Emmert-Streib, F. & Glazko, G.V. Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Computational Biology* **7**, e1002053 (2011).
97. Song, S. & Black, M.A. Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics* **9**, 502 (2008).
98. Hedegaard, J. et al. Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. *BMC Proc* **3 Suppl 4**, S5 (2009).
99. Mazumder, A., Palma, A.J. & Wang, Y. Validation and integration of gene-expression signatures in cancer. *Expert Rev Mol Diagn* **8**, 125-128 (2008).
100. Cary, M.P., Bader, G.D. & Sander, C. Pathway information for systems biology. *FEBS Lett* **579**, 1815-1820 (2005).
101. Tsui, I.F., Chari, R., Buys, T.P. & Lam, W.L. Public databases and software for the pathway analysis of cancer genomes. *Cancer Inform* **3**, 379-397 (2007).
102. Nam, D. & Kim, S.Y. Gene-set approach for expression pattern analysis. *Brief Bioinform* **9**, 189-197 (2008).
103. Thomas, D.C. et al. Use of pathway information in molecular epidemiology. *Hum Genomics* **4**, 21-42 (2009).
104. Vaske, C.J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245 (2010).
105. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
106. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-114 (2012).
107. Nishimura, D. BioCarta. *Biotech Software Internet Report* **2**, 117-120 (2001).
108. Schaefer, C.F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-679 (2009).

109. Hastie, T. & Stuetzle, W. Principal Curves. *J Am Stat Assoc* **84**, 502-516 (1989).
110. Sheffer, M. et al. Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. *Proceedings of the National Academy of Sciences* **106**, 7131-7136 (2009).
111. Lee, J.C. et al. Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS Med* **3**, e485 (2006).
112. Verhaak, R.G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).
113. Woerner, B.M. et al. Suppression of G-protein-coupled receptor kinase 3 expression is a feature of classical GBM that is required for maximal growth. *Mol Cancer Res* **10**, 156-166 (2012).
114. Phillips, H.S. et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173 (2006).
115. National Cancer Institute. REMBRANDT home page. <*http://rembrandt.nci.nih.gov*> **Release 07-27-2010** (2005).
116. Madhavan, S. et al. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res* **7**, 157-167 (2009).
117. Peto, R. & Peto, J. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General)* **135**, 185 (1972).
118. Demuth, T. et al. MAP-ing glioma invasion: mitogen-activated protein kinase kinase 3 and p38 drive glioma invasion and progression and predict patient survival. *Mol Cancer Ther* **6**, 1212-1222 (2007).
119. Mawrin, C. et al. Prognostic relevance of MAPK expression in glioblastoma multiforme. *Int J Oncol* **23**, 641-648 (2003).
120. Glassmann, A. et al. Pharmacological targeting of the constitutively activated MEK/MAPK-dependent signaling pathway in glioma cells inhibits cell proliferation and migration. *Int J Oncol* **39**, 1567-1575 (2011).
121. Mason, W.P. et al. A phase II study of the Ras-MAPK signaling pathway inhibitor TLN-4601 in patients with glioblastoma at first progression. *J Neurooncol* **107**, 343-349 (2012).
122. Krakstad, C. & Chekenya, M. Survival signalling and apoptosis resistance in glioblastomas: opportunities for targeted therapeutics. *Mol Cancer* **9**, 135 (2010).

123. Shaw, R.J. & Cantley, L.C. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature* **441**, 424-430 (2006).
124. Newton, H.B. Molecular neuro-oncology and development of targeted therapeutic strategies for brain tumors. Part 2: PI3K/Akt/PTEN, mTOR, SHH/PTCH and angiogenesis. *Expert Rev Anticancer Ther* **4**, 105-128 (2004).
125. Hildebrandt, M.A. et al. Genetic variations in the PI3K/PTEN/AKT/mTOR pathway are associated with clinical outcomes in esophageal cancer patients treated with chemoradiotherapy. *J Clin Oncol* **27**, 857-871 (2009).
126. Krause, D.S. & Van Etten, R.A. Tyrosine kinases as targets for cancer therapy. *N Engl J Med* **353**, 172-187 (2005).
127. Andersson, U. et al. Epidermal growth factor receptor family (EGFR, ErbB2-4) in gliomas and meningiomas. *Acta Neuropathol* **108**, 135-142 (2004).
128. Dong, Y. et al. Selective inhibition of PDGFR by imatinib elicits the sustained activation of ERK and downstream receptor signaling in malignant glioma cells. *Int J Oncol* **38**, 555-569 (2011).
129. Desgrosellier, J.S. & Cheresh, D.A. Integrins in cancer: biological implications and therapeutic opportunities. *Nat Rev Cancer* **10**, 9-22 (2010).
130. Rascher, G. et al. Extracellular matrix and the blood-brain barrier in glioblastoma multiforme: spatial segregation of tenascin and agrin. *Acta Neuropathol* **104**, 85-91 (2002).
131. Thapar, K. et al. Overexpression of the growth-hormone-releasing hormone gene in acromegaly-associated pituitary tumors. An event associated with neoplastic progression and aggressive behavior. *Am J Pathol* **151**, 769-784 (1997).
132. Ara, T. & DeClerck, Y.A. Mechanisms of invasion and metastasis in human neuroblastoma. *Cancer Metastasis Rev* **25**, 645-657 (2006).
133. Penson, R.T. et al. Cytokines IL-1beta, IL-2, IL-6, IL-8, MCP-1, GM-CSF and TNFalpha in patients with epithelial ovarian cancer and their relationship to treatment with paclitaxel. *Int J Gynecol Cancer* **10**, 33-41 (2000).
134. van Rossum, A.P. et al. Granulocytosis and thrombocytosis in renal cell carcinoma: a pro-inflammatory cytokine response originating in the tumour. *Neth J Med* **67**, 191-194 (2009).
135. Eynard, A.R. Potential of essential fatty acids as natural therapeutic products for human tumors. *Nutrition* **19**, 386-388 (2003).
136. Nigro, E.A. et al. Antitumor IgE adjuvanticity: key role of Fc epsilon RI. *J Immunol* **183**, 4530-4536 (2009).

137. Hauck, C.R., Hsia, D.A. & Schlaepfer, D.D. The focal adhesion kinase--a regulator of cell migration and invasion. *IUBMB Life* **53**, 115-119 (2002).
138. van Groeninghen, J.C., Kiesel, L., Winkler, D. & Zwirner, M. Effects of luteinising-hormone-releasing hormone on nervous-system tumours. *Lancet* **352**, 372-373 (1998).
139. Cook, T. & Sheridan, W.P. Development of GnRH antagonists for prostate cancer: new approaches to treatment. *Oncologist* **5**, 162-168 (2000).
140. Marelli, M.M. et al. Novel insights into GnRH receptor activity: role in the control of human glioblastoma cell proliferation. *Oncol Rep* **21**, 1277-1282 (2009).
141. Park, D.W., Choi, K.C., MacCalman, C.D. & Leung, P.C. Gonadotropin-releasing hormone (GnRH)-I and GnRH-II induce cell growth inhibition in human endometrial cancer cells: involvement of integrin beta3 and focal adhesion kinase. *Reprod Biol Endocrinol* **7**, 81 (2009).
142. Raghu, H., Gondi, C.S., Dinh, D.H., Gujrati, M. & Rao, J.S. Specific knockdown of uPA/uPAR attenuates invasion in glioblastoma cells and xenografts by inhibition of cleavage and trafficking of Notch -1 receptor. *Mol Cancer* **10**, 130 (2011).
143. Pawlak, K., Ulazka, B., Mysliwiec, M. & Pawlak, D. Vascular endothelial growth factor and uPA/suPAR system in early and advanced chronic kidney disease patients: a new link between angiogenesis and hyperfibrinolysis? *Transl Res* (2012).
144. Wen, P.Y., Lee, E.Q., Reardon, D.A., Ligon, K.L. & Alfred Yung, W.K. Current clinical development of PI3K pathway inhibitors in glioblastoma. *Neuro Oncol* **14**, 819-829 (2012).
145. Waugh, M.G. Phosphatidylinositol 4-kinases, phosphatidylinositol 4-phosphate and cancer. *Cancer Lett* (2012).
146. Jamieson, S. et al. A drug targeting only p110alpha can block phosphoinositide 3-kinase signalling and tumour growth in certain cell types. *Biochem J* **438**, 53-62 (2011).
147. Nakata, S. et al. LGR5 is a marker of poor prognosis in glioblastoma and is required for survival of brain cancer stem-like cells. *Brain Pathol* (2012).
148. Rossi, M. et al. beta-catenin and Gli1 are prognostic markers in glioblastoma. *Cancer Biol Ther* **11**, 753-761 (2011).
149. Eyler, C.E. & Rich, J.N. Survival of the fittest: cancer stem cells in therapeutic resistance and angiogenesis. *J Clin Oncol* **26**, 2839-2845 (2008).
150. Behrens, J. et al. Loss of epithelial differentiation and gain of invasiveness correlates with tyrosine phosphorylation of the E-

cadherin/beta-catenin complex in cells transformed with a temperature-sensitive v-SRC gene. *J Cell Biol* **120**, 757-766 (1993).
151. Aikawa, T. et al. Glypican-1 modulates the angiogenic and metastatic potential of human and mouse cancer cells. *J Clin Invest* **118**, 89-99 (2008).
152. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
153. Bertrand, J. et al. Cancer stem cells from human glioma cell line are resistant to Fas-induced apoptosis. *Int J Oncol* **34**, 717-727 (2009).
154. Calzada, M.J. et al. Alpha4beta1 integrin mediates selective endothelial cell responses to thrombospondins 1 and 2 in vitro and modulates angiogenesis in vivo. *Circ Res* **94**, 462-470 (2004).
155. Aoudjit, F. & Vuori, K. Integrin signaling in cancer cell survival and chemoresistance. *Chemother Res Pract* **2012**, 283181 (2012).
156. Erikstein, B.K. et al. Expression of CD18 (integrin beta 2 chain) correlates with prognosis in malignant B cell lymphomas. *Br J Haematol* **83**, 392-398 (1993).
157. Terol, M.J. et al. Expression of beta-integrin adhesion molecules in non-Hodgkin's lymphoma: correlation with clinical and evolutive features. *J Clin Oncol* **17**, 1869-1875 (1999).
158. Lathia, J.D. et al. Integrin alpha 6 regulates glioblastoma stem cells. *Cell Stem Cell* **6**, 421-432 (2010).
159. Velpula, K.K. et al. Glioma stem cell invasion through regulation of the interconnected ERK, integrin alpha6 and N-cadherin signaling pathway. *Cell Signal* (2012).
160. Friedrichs, K. et al. High expression level of alpha 6 integrin in human breast carcinoma is correlated with reduced survival. *Cancer Res* **55**, 901-906 (1995).
161. Akasaki, Y. et al. A peroxisome proliferator-activated receptor-gamma agonist, troglitazone, facilitates caspase-8 and -9 activities by increasing the enzymatic activity of protein-tyrosine phosphatase-1B on human glioma cells. *J Biol Chem* **281**, 6165-6174 (2006).
162. Evangelisti, C. et al. MiR-128 up-regulation inhibits Reelin and DCX expression and reduces neuroblastoma cell motility and invasiveness. *FASEB J* **23**, 4276-4287 (2009).
163. Becker, J., Frohlich, J., Perske, C., Pavlakovic, H. & Wilting, J. Reelin signalling in neuroblastoma: Migratory switch in metastatic stages. *Int J Oncol* **41**, 681-689 (2012).
164. Brule, S. et al. Glycosaminoglycans and syndecan-4 are involved in SDF-1/CXCL12-mediated invasion of human epitheloid

carcinoma HeLa cells. *Biochim Biophys Acta* **1790**, 1643-1650 (2009).
165. Huang, W., Chiquet-Ehrismann, R., Moyano, J.V., Garcia-Pardo, A. & Orend, G. Interference of tenascin-C with syndecan-4 binding to fibronectin blocks cell adhesion and stimulates tumor cell proliferation. *Cancer Res* **61**, 8586-8594 (2001).
166. Ridgway, L.D., Wetzel, M.D. & Marchetti, D. Modulation of GEF-H1 induced signaling by heparanase in brain metastatic melanoma cells. *J Cell Biochem* **111**, 1299-1309 (2010).
167. Naganuma, H. et al. Quantification of thrombospondin-1 secretion and expression of alphavbeta3 and alpha3beta1 integrins and syndecan-1 as cell-surface receptors for thrombospondin-1 in malignant glioma cells. *J Neurooncol* **70**, 309-317 (2004).
168. Xu, Y., Yuan, J., Zhang, Z., Lin, L. & Xu, S. Syndecan-1 expression in human glioma is correlated with advanced tumor progression and poor prognosis. *Mol Biol Rep* (2012).
169. Tibshirani, R. Principal curves revisited. *Stat Comput* **2**, 183-190 (1992).
170. Leblanc, M. & Tibshirani, R. Adaptive Principal Surfaces. *J Am Stat Assoc* **89**, 53-64 (1994).
171. Bishop, C.M., Svensen, M. & Williams, C.K.I. GTM: The generative topographic mapping. *Neural Comput* **10**, 215-234 (1998).
172. Kegl, B., Krzyzak, A., Linder, T. & Zeger, K. Learning and design of principal curves. *Ieee T Pattern Anal* **22**, 281-297 (2000).
173. Delicado, P. Another look at principal curves and surfaces. *J Multivariate Anal* **77**, 84-116 (2001).
174. Lawrence, N. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J Mach Learn Res* **6**, 1783-1816 (2005).
175. Urbach, S. in Physics of Complex System, Vol. MSc (MSc Thesis, Weizmann Institute of Science, Rehovot; 2006).
176. Ballman, K.V., Grill, D.E., Oberg, A.L. & Therneau, T.M. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* **20**, 2778-2786 (2004).
177. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
178. Aaronson, P.I., Ward, J., Wiener, C.M., Schulman, S.P. & Gill, J.S. The Cardiovascular System at a Glance. (Blackwell Science, 2000).

179. Finn, A.V., Nakano, M., Narula, J., Kolodgie, F.D. & Virmani, R. Concept of vulnerable/unstable plaque. *Arterioscler Thromb Vasc Biol* **30**, 1282-1292 (2010).
180. Falk, E., Shah, P.K. & Fuster, V. Coronary plaque disruption. *Circulation* **92**, 657-671 (1995).
181. Burke, A.P. et al. Coronary risk factors and plaque morphology in men with coronary disease who died suddenly. *N Engl J Med* **336**, 1276-1282 (1997).
182. Ambrose, J.A. & Fuster, V. The risk of coronary occlusion is not proportional to the prior severity of coronary stenoses. *Heart (British Cardiac Society)* **79**, 3-4 (1998).
183. Mann, J.M. & Davies, M.J. Vulnerable plaque. Relation of characteristics to degree of stenosis in human coronary arteries. *Circulation* **94**, 928-931 (1996).
184. Topol, E.J. The genetics of heart attack. *Heart (British Cardiac Society)* **92**, 855-861 (2006).
185. Patino, W.D. et al. Circulating transcriptome reveals markers of atherosclerosis. *Proc Natl Acad Sci U S A* **102**, 3423-3428 (2005).
186. Wingrove, J.A. et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. *Circ Cardiovasc Genet* **1**, 31-38 (2008).
187. Sinnaeve, P.R. et al. Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One* **4**, e7037 (2009).
188. Rosenberg, S. et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. *Ann Intern Med* **153**, 425-434 (2010).
189. Goldstein, J.A. et al. Multiple complex coronary plaques in patients with acute myocardial infarction. *N Engl J Med* **343**, 915-922 (2000).
190. Qiao, J.H. & Fishbein, M.C. The severity of coronary atherosclerosis at sites of plaque rupture with occlusive thrombosis. *Journal of the American College of Cardiology* **17**, 1138-1142 (1991).
191. Rehr, R., Disciascio, G., Vetrovec, G. & Cowley, M. Angiographic morphology of coronary artery stenoses in prolonged rest angina: evidence of intracoronary thrombosis. *Journal of the American College of Cardiology* **14**, 1429-1437 (1989).
192. Kerner, A. et al. Relation of C-reactive protein to coronary collaterals in patients with stable angina pectoris and coronary artery disease. *The American journal of cardiology* **99**, 509-512 (2007).

193. Gensini, G.G. A more meaningful scoring system for determining the severity of coronary heart disease. *Am J Cardiol* **51**, 606 (1983).
194. Montorsi, P. et al. Association between erectile dysfunction and coronary artery disease. Role of coronary clinical presentation and extent of coronary vessels involvement: the COBRA trial. *European heart journal* **27**, 2632-2639 (2006).
195. James, S. et al. An acute inflammatory reaction induced by myocardial damage is superimposed on a chronic inflammation in unstable coronary artery disease. *Am Heart J* **149**, 619-626 (2005).
196. Ridker, P.M. et al. C-reactive protein levels and outcomes after statin therapy. *The New England journal of medicine* **352**, 20-28 (2005).
197. Kuhn, K. et al. A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res* **14**, 2347-2356 (2004).
198. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408 (2001).
199. Lin, S.M., Du, P., Huber, W. & Kibbe, W.A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* **36**, e11 (2008).
200. Benito, M. et al. Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105-114 (2004).
201. Archer, K.J. & Reese, S.E. Detection call algorithms for high-throughput gene expression microarray data. *Brief Bioinform* **11**, 244-252 (2010).
202. Tsafrir, D. et al. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* **21**, 2301-2308 (2005).
203. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
204. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
205. Aronson, D. et al. Effect of obesity on the relationship between plasma C-reactive protein and coronary artery stenosis in patients with stable angina. *Atherosclerosis* **185**, 137-142 (2006).
206. Khera, A. et al. Relationship between C-reactive protein and subclinical atherosclerosis: the Dallas Heart Study. *Circulation* **113**, 38-43 (2006).

207. McCaffrey, T.A. et al. High-level expression of Egr-1 and Egr-1-inducible genes in mouse and human atherosclerosis. *The Journal of clinical investigation* **105**, 653-662 (2000).
208. Khachigian, L.M. Early growth response-1 in cardiovascular pathobiology. *Circulation research* **98**, 186-191 (2006).
209. Goetze, S. et al. TNFalpha induces expression of transcription factors c-fos, Egr-1, and Ets-1 in vascular lesions through extracellular signal-regulated kinases 1/2. *Atherosclerosis* **159**, 93-101 (2001).
210. Galea, J. et al. Interleukin-1 beta in coronary arteries of patients with ischemic heart disease. *Arteriosclerosis, thrombosis, and vascular biology* **16**, 1000-1006 (1996).
211. Tipping, P.G. & Hancock, W.W. Production of tumor necrosis factor and interleukin-1 by macrophages from human atheromatous plaques. *Am J Pathol* **142**, 1721-1728 (1993).
212. Albasanz-Puig, A. et al. Oncostatin M is expressed in atherosclerotic lesions: A role for Oncostatin M in the pathogenesis of atherosclerosis. *Atherosclerosis* **216**, 292-298 (2011).
213. Demyanets, S. et al. Oncostatin M-enhanced vascular endothelial growth factor expression in human vascular smooth muscle cells involves PI3K-, p38 MAPK-, Erk1/2- and STAT1/STAT3-dependent pathways and is attenuated by interferon-gamma. *Basic Res Cardiol* **106**, 217-231 (2011).
214. Modur, V. et al. Oncostatin M is a proinflammatory mediator. In vivo effects correlate with endothelial cell expression of inflammatory cytokines and adhesion molecules. *The Journal of clinical investigation* **100**, 158-168 (1997).
215. Belton, O., Byrne, D., Kearney, D., Leahy, A. & Fitzgerald, D.J. Cyclooxygenase-1 and -2-dependent prostacyclin formation in patients with atherosclerosis. *Circulation* **102**, 840-845 (2000).
216. Baker, C.S. et al. Cyclooxygenase-2 is widely expressed in atherosclerotic lesions affecting native and transplanted human coronary arteries and colocalizes with inducible nitric oxide synthase and nitrotyrosine particularly in macrophages. *Arteriosclerosis, thrombosis, and vascular biology* **19**, 646-655 (1999).
217. Zhan, Y. et al. Effects of dominant-negative c-Jun on platelet-derived growth factor-induced vascular smooth muscle cell proliferation. *Arterioscler Thromb Vasc Biol* **22**, 82-88 (2002).
218. Ahmad, M., Theofanidis, P. & Medford, R.M. Role of activating protein-1 in the regulation of the vascular cell adhesion molecule-1 gene expression by tumor necrosis factor-alpha. *J Biol Chem* **273**, 4616-4621 (1998).

219. Bavendiek, U. et al. Induction of tissue factor expression in human endothelial cells by CD40 ligand is mediated via activator protein 1, nuclear factor kappa B, and Egr-1. *J Biol Chem* **277**, 25032-25039 (2002).
220. Wu, G.S. et al. KILLER/DR5 is a DNA damage-inducible p53-regulated death receptor gene. *Nature genetics* **17**, 141-143 (1997).
221. Stawowczyk, M., Van Scoy, S., Kumar, K.P. & Reich, N.C. The interferon stimulated gene 54 promotes apoptosis. *J Biol Chem* **286**, 7257-7266 (2011).
222. Michowitz, Y. et al. The involvement of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) in atherosclerosis. *Journal of the American College of Cardiology* **45**, 1018-1024 (2005).
223. Libby, P. & Crea, F. Clinical implications of inflammation for cardiovascular primary prevention. *European heart journal* **31**, 777-783 (2010).
224. Mauriello, A. et al. Diffuse and active inflammation occurs in both vulnerable and stable plaques of the entire coronary tree: a histopathologic study of patients dying of acute myocardial infarction. *Journal of the American College of Cardiology* **45**, 1585-1593 (2005).
225. Buffon, A. et al. Widespread coronary inflammation in unstable angina. *N Engl J Med* **347**, 5-12. (2002).

11.Declaration

The work on the detection and analysis of somatic rearrangements was done in collaboration with Dr. Gad Getz, Director of Cancer Genome Computational Analysis at the Broad Institute of MIT and Harvard and his team. My main contribution to this effort was the development of BreakPointer, the tool pinpointing the rearrangement breakpoint, and the pan-cancer analysis of somatic rearrangements, described in chapter 4, all done entirely by me. Also, I have contributed equally with Dr. Mike Berger, Dr. Mike Lawrence and Dr. Francesca Demichelis to the prostate cancer study, and helped with the rearrangement analysis of other types of cancer, as described in chapter 5.

I have developed Pathifier- the method to quantify pathway deregulation, and applied it to analyze glioblastoma, as described in chapter 6, and helped Dr. Michal Sheffer in the appling Pathifier to colorectal cancer and analyzing its results (not described in this thesis, but appearing in our publication #2).

The work on the double EGF pulse was done in collaboration with prof. Yosef Yarden's lab that did all the experimental work, while I have done all the bioinformatic analysis, in close collaboration with Dr. Yaara Zwang and the help of Dr. Tal Shay's insights.

The work on acute myocardial infarction was done in collaboration with Dr. Doron Aronson and Sagi Nahum of the Technion, who did all the experimental work, while I have done the bioinformatic analysis.

OPEN ACCESS Freely available online

PLoS one

Do Two Machine-Learning Based Prognostic Signatures for Breast Cancer Capture the Same Biological Processes?

Yotam Drier, Eytan Domany*

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel

Abstract

The fact that there is very little if any overlap between the genes of different prognostic signatures for early-discovery breast cancer is well documented. The reasons for this apparent discrepancy have been explained by the limits of simple machine-learning identification and ranking techniques, and the biological relevance and meaning of the prognostic gene lists was questioned. Subsequently, proponents of the prognostic gene lists claimed that different lists do capture similar underlying biological processes and pathways. The present study places under scrutiny the validity of this claim, for two important gene lists that are at the focus of current large-scale validation efforts. We performed careful enrichment analysis, controlling the effects of multiple testing in a manner which takes into account the nested dependent structure of gene ontologies. In contradiction to several previous publications, we find that the only biological process or pathway for which statistically significant concordance can be claimed is cell proliferation, a process whose relevance and prognostic value was well known long before gene expression profiling. We found that the claims reported by others, of wider concordance between the biological processes captured by the two prognostic signatures studied, were found either to be lacking statistical rigor or were in fact based on addressing some other question.

Citation: Drier Y, Domany E (2011) Do Two Machine-Learning Based Prognostic Signatures for Breast Cancer Capture the Same Biological Processes? PLoS ONE 6(3): e17795. doi:10.1371/journal.pone.0017795

Editor: Wael El-Rifai, Vanderbilt University Medical Center, United States of America

Received October 31, 2010; **Accepted** February 14, 2011; **Published** March 14, 2011

Funding: This research was supported by the Leir Charitable Foundation, a Weizmann-Mario Negri collaborative research grant and by a grant from the German Research Foundation (DIP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: eytan.domany@weizmann.ac.il

Introduction

Technological advances made during the last decade have allowed measurement of enormous amounts of molecular data from a tumor tissue resected from a particular subject. The main challenge of modern cancer research is bridging the gap between these data and clinically significant questions that need urgent answers, such as prognosis and prediction of response to therapy.

The first issue, of prognosis, is highly relevant, since it is used to decide whether to subject a patient to chemotherapy. This decision is extremely important for the individual as well as for society for three main reasons. First, nearly all available chemotherapy is detrimental to the patient, since it adversely affects healthy tissue as well as the malignant one, at which it is aimed. Second, some of the side effects, even if they do not have a direct impact on the patient's physical well-being, may cause considerable psychological damage and hardship. Finally – chemotherapy is extremely expensive.

It is well known that for many cancers prognosis and the need for therapy may vary widely; while in some cases surgery and adjuvant radiotherapy suffice to eradicate the disease, other tumors are very aggressive, will recur, metastase and kill the patient. While aggressive tumors call for chemotherapy, overtreatment of "good outcome" patients by administering unneeded chemotherapy is, unfortunately, very common. This is the case particularly in breast cancer, where increased awareness has brought, through regular frequent checkups, a considerable increase in the number of early discovery cases, of small tumors of low stage and grade.

It is believed that the currently accepted clinical-pathological criteria for administering chemotherapy gives rise to overtreatment of a very large fraction of early discovery breast cancer patients. Therefore, there is an acute need for reliable biomarkers that can, on the basis of measurements done on the primary tumor tissue, differentiate poor from good outcome.

A large number of methods have been introduced to generate biomarkers from available molecular information (in particular from gene expression microarray data – see [1,2,3,4] for reviews). Two prognostic platforms based on expression signatures are commercially available: OncotypeDx, based on a 21-gene signature measured on paraffin-embedded samples by polymerase chain reaction (PCR) [5], and Mammaprint, the 70-gene "Amsterdam signature" measured by a microarray [6,7,8,9].

Considerable criticism has been raised about the following aspects of several proposed signatures: lack of robustness, various statistical and machine-learning related problems, low success rates for the cases that are hard to prognosticate by existing methods, and lack of biological meaning of gene lists, that were obtained without biological guidance.

The first criticism, concerning the statistical validity and robustness of the reported gene lists, focuses on the fact that in many cases the reported signatures were derived and tested in only one particular way, which was arbitrarily selected out of many equally legitimate ones. For example, one can split the samples into a training set and a test set in a combinatorially large number of ways. Hence the entire analysis, including training, gene selection and testing, can be repeated many times, using the same data, but splitting differently the samples into training and test sets.

Each such split can be viewed as a particular instance of the analysis, and by performing many such repeats, one can generate distributions of various quantities of interest. In particular, one can calculate for each split the success rate, defined as the fraction of successful predictions of outcome on the test set, and estimate the distribution of the success rates by repeating the analysis many times. Once this distribution is known, one can estimate the probability to find a success rate as good as, or better, than the one reported in the actual published study (for which the analysis was repeated). When this was done [10,11], the results of many studies have been demonstrated to be "overoptimistic" [11]; the success rate that was actually reported had a much lower than acceptable probability of being observed. The overoptimistic reported success rates of many studies were explained by falling into various statistical pitfalls [2,12]. These include severe overtraining [2], due mainly to "information leak" which has been explicitly identified in a number of cases [2,13]. The term information leak refers to allowing usage of any information about the test set during the training phase. Another issue concerns the *prognostic lists* of genes (which are the ones that are actually placed on a prognostic device [7]). The genes that appear in the prognostic list of a particular study were selected by ranking all the tested genes (for example, on the basis of the correlation of their expression values (measured over the training samples) with outcome. These lists were shown to lack robustness [10] for the sample sizes used [14,15]; i.e. the prognostic gene lists changed almost completely when the procedure was repeated. It has been shown [14,15] that if a training set of ~100 early breast cancer samples is used to rank ~10,000 genes (by their correlation with outcome), and the ~100 top genes are selected as the prognostic set, repeating the procedure (with a different set of training samples) will produce a new gene list, whose overlap with the first one is typically 2–3%. Since the different gene lists obtained even from the same particular study are very unstable against repeating the analysis, one clearly expects even less overlap between lists produced by different studies (in which different patients, different microarray facilities and even different platforms were used). In response to this criticism it was stated that if two divergent lists provide concordant prognostication and acceptable success rates, one should not care about their lack of robustness [16]. This response was countered, however, by criticism raised against the criteria that were used to assess the success rates of several expression-based classifiers [17], and various publications questioned whether they actually performed better than either a single-gene based classifier [18] or one that uses classical clinical and pathological indicators [19,20]. The issue of concordance [21] or lack thereof [17,22] between different prognostic signatures was also debated.

The points of criticism described above address either technical issues that concern the standard machine-learning approaches taken by most derivations of prognostic signatures, or the clinical utility of the resulting classifiers. In the present study we focus on a third type of criticism, directed at the lack of biological meaning of various prognostic signatures. Some signatures [5,23,24,25,26,27] did use biological and clinical knowledge to assemble their predictive genes. We did not consider the Oncotype DX recurrence signature [28], which was constructed by carefully picking genes from relevant pathways (and therefore indeed, capture many pathways); the P53 signature [24], BMI1 signature [27] and wound response signature [23,29], each of which was constructed to capture a specific pathway, as their names suggest (and therefore indeed mainly capture the desired pathway); or the genomic grade signature [30] that was constructed specifically to capture histological grade (and was found to include mostly proliferation-related genes [31]). Our focus is on prognostic gene lists which were derived using the "top-down" approach as defined in [32], that is, either using no biological guidance at all for feature selection and training – e.g. the Amsterdam signature [7], or using very minimal biological input, such as for the 76-gene Rotterdam signature [33], which treated ER+ and ER− breast cancers separately.

According to the critics, these prognostic gene lists lack clear biological interpretation and probably contain no biologically relevant discovery. In response to this criticism it was claimed by some [34] that the biological processes that were represented by the activities of the genes on such divergent lists did, in fact, exhibit considerable similarity. If correct, this claim gives one more reason why one should not worry about the fact that the gene lists of different studies had no overlap; furthermore, this would also answer the criticism regarding biological meaning.

The claim that divergent gene lists from different studies do reflect the activities of similar cancer-related pathways and biological processes seems to be advocated and accepted by many [1,9,15,21,34,35,36]. Only a few studies [37,38,39,40] have, however, actually tried to substantiate these claims in a quantitative manner. The aim of our study is to test the validity of these claims in a way which we believe is conceptually and statistically sound.

In what follows, we first present the guiding principles that must be adhered to in order to test properly these claims, and then we review critically the studies mentioned above. Next we present our results obtained when the analysis is carried out for two important signatures [7,33] according to our guiding principles. We conclude that the only biological processes and pathways that are significantly represented by both these signatures are cell proliferation and its variants.

The guiding principles of the present study

Our aim here was to test critically the claims that two different machine-learning based prognostic gene lists capture similar biological processes. To this end we examined the two most established outcome prediction signatures, the 70 gene list of van't Veer et al. [7] and the lists defined by Wang et al. [33], both the 60 gene ER+ signature and the complete 76-gene list. We have chosen these two signatures as they were learned independently and without forcing specific biological pathway-based knowledge.

We adopted the following guiding principles in designing our test:

1. Use only the genes that actually appear in the prognostic lists.
2. Identify over-represented biological processes by means of enrichment analysis.
3. Address the problem of false discoveries generated by multiple comparisons that are made, but take into account all the dependencies and nested structures present in the ontologies used.
4. Use more than one gene ontology, to minimize dependence on incomplete or deficient class assignments.

The rationale for the first principle is the following. As stated above, our aim is to test, in a statistically correct way, the claim that was voiced by proponents of the proposed prognostic lists, that different lists do capture the same biological processes. To test this claim, one is not supposed to use larger gene lists, which could have been derived from the same experiment by some other means. We are neither claiming that gene expression cannot possibly capture important and biologically relevant prognostic information, nor are we attempting to demonstrate how one could, in principle, capture such information.

In fact it is likely that the full data gathered in these studies do reflect similar deregulation of a few common relevant pathways, but it remains to be proven that this similarity is captured in the actual proposed gene lists. In that regard it is worth mentioning that when standard machine-learning methods are used to select features (genes) for a classifier, the number of selected features cannot exceed significantly the number of samples that happened to be available for training [41] (at the time when the study was first performed and the gene lists were selected). Otherwise the classifier is trained to recognize the noise in the particular training set used, and will fail on any test set (since while the true "signal" is the same in the training and test sets, the noise is completely independent). This limitation might restrict the selected number of genes and produce lists of selected genes that are too short to capture the necessary biological processes. Two possible ways to overcome this are producing much longer gene lists (for which much more training samples must become available), or use biologically relevant knowledge based considerations to select the predictive genes.

The second principle states that a generally accepted method [42] be used to assess enrichment of a pathway or biological process by the prognostic list.

The third principle – the necessity for taking into consideration the false discoveries [43,44] that arise when multiple comparisons are made - cannot be overemphasized [41]. A problem arises when one performs enrichment analysis of GO (gene ontology) terms [45], such as Biological Processes (GOBP). When the number of GO terms is taken as the number of independent tests, it is likely that not a single term will pass any of the available procedures [46] that control the FDR. The reason is that because of the nested and overlapping structure of the ontology, the many terms tested are not independent and hence the standard methods that control the FDR are much too stringent [47,48] (to understand this point, imagine that in fact we have one single term which for some reason is repeated 1000 times – while only a single test was performed, naively we may think that 1000 hypotheses were tested). The trivial resolution of this problem, of ignoring multiple testing altogether and make no attempt to control the FDR, goes to the opposite extreme and is way too permissive, generating a very large number of false positive apparently enriched GO terms.

We present and compare three ways to deal with the problem of multiple comparisons. The first is to apply the standard Benjamini-Hochberg procedure [43] to control the FDR, ignoring the nested structure of the ontologies. We show that this procedure, which is probably too stringent, finds almost no commonly enriched biological processes or pathways. The second and third are two different ways, explained in detail in the Methods section, designed to deal with multiple comparisons while taking the dependencies and nested structure of the ontologies into account.

The fourth principle stems from the known fact that ontologies are far from being perfect, and probably contain some incorrectly assigned genes; testing a claimed enrichment for more than one ontology or database is prudent.

The manner in which each of these points is implemented is explained in detail in the Methods section below.

Brief review of previous work

The abstract of Yu et al. [40] states that "We show that divergent gene sets classifying patients for the same clinical endpoint represent similar biological processes …". They addressed this issue indirectly by using expression data of 344 early discovery breast cancer patients; the same analysis was done separately for the ER+ and ER− cases. 80 samples were selected at random as training set; Cox regression analysis was performed to identify the 100 genes whose expression was most correlated with distant metastasis-free survival time. These "top 100" genes were analyzed for enrichment of 304 GOBP (selected, using some arbitrary thresholds, from the total list of GOBP). The enrichment analysis was done as follows: hypergeometric p-values were calculated (Fisher's exact test) for over-representation of the genes that belong to a GOBP among the 100 "top genes", and if the number of genes exceeded one and the p-value was less than 0.05, the GOBP was declared enriched. No correction for multiple comparisons (of either genes or GOPB) was used, and no special treatment to the GOBP dependence (due genes that appear in several GOBPs) was offered. This analysis was repeated 500 times, yielding 500 lists of enriched GOBP. The 20 GOBP that had the highest number of appearances were assembled, for ER+ and ER−, yielding 36 "core pathways" (4 appeared on both lists). Finally, several published prognostic gene lists were analyzed for enrichment among the 304 GOBP and among the 36 core pathways, and using the hypergeometric distribution, significant overrepresentation of the core pathways was reported.

This analysis is too permissive mainly because no FDR correction for multiple comparisons was used at all. Moreover, several arbitrary and unjustified thresholds were used for selection of GOBP to be tested and for identification of enriched GOBP; the sets of enriched GOBPs obtained for each pair of prognostic gene lists were not compared directly, but each was compared to the list of core pathways defined above; only one database of biological pathways and processes was used for the study.

Shen et al. [38] have followed similar guidelines to those we suggest. They actually don't find a statistically significant number of pathways common to the Wang and van't Veer lists (this fact is not emphasized, but see Figure 1 of their paper). Moreover, the statistical significance of the overlaps they report is due to an unusual definition of the p-value. Namely, if they find that the tested prognostic list contains k genes from a pathway, they estimate the p-value as the probability that a random gene list will contain more than k genes from the pathway- $p(x>k)$, instead of using the standard definition, i.e. the probability to find k or more than k such genes- $p(x\geq k)$. These two probabilities are nearly the same for most situations, but can be quite different when the list is very short (small k), as is the case here, where often $k = 1$. Table 4 in their paper shows what appears to be significant overlap between several signatures, but in fact there is only one single gene of the 70 gene list that belongs to each of the 'enriched' pathways. Given that 50 genes from the 70 are annotated, chosen out of 11342 genes on the chip (the "population"), and that, for example, the RECK pathway (one of the five presented as significantly over-represented and shared in Table 4) has 8 genes from the population, a naïve hypergeometric test will conclude a p-value of 0.035, while Shen's measure will indicate a much higher significance, of 5.24*10-4. Checking the hypothesis for all probes (not just annotated ones) will increase the p-value further. The naïve hypergeometric high p-values will not pass a reasonable FDR on the 552 hypotheses checked. The other 4 pathways also have only one gene among the 50, and since these pathways contain more genes than RECK, their p-values will only be bigger. Even if one chooses to ignore the 70 gene list, and look for pathways common only to the three other signatures checked in the paper, only the breast cancer estrogen signaling pathway is found to be over represented in all. Repeating this analysis using the standard definition of the p-value, we found that for the 70 gene list no pathway passes at any reasonable FDR, and even if we ignored the 70 gene list, still only the breast cancer estrogen signaling pathway was over represented in all the other three signatures tested.

Reyal et al. [37] have not approached the question of pathway convergence of the signatures directly, but instead aimed at offering a new, pathway based predictor. In order to do so they have used a large number of tumor expression profiles measured by the Affymetrix 133A platform, started from seven published signatures and used them to create enlarged signatures. These contained all the genes correlated to the original signature, revealing large gene clusters that differentiated good from bad outcome. A careful pathway analysis discovered common pathways which were then used to build new, more promising predictors. In our context, however, one must be careful not to deduce from this study that there is biological agreement between the actual seven signatures they studied, as their analysis was done on highly enlarged gene lists.

Sole et al. [39] have tested different signatures of different cancers, including breast cancer signatures [5,29,49,50,51,52], by two main approaches. The first was to check for overrepresentation of transcription factor targets as predicted by motif analysis and chip experiments. The second was to check on a few datasets for correlations between the signature genes and the various pathway genes. The first approach identified targets of E2F and ER, as well as cell cycle genes, to be common to many of the signatures. Note that E2F is a major proliferation regulator and many of its targets correlate with proliferation rate. They have raised also the possibility that AHR, MYB and MYC targets are overrepresented in a few of the signatures. The second approach identified mitosis and possibly immune response as related to some of the breast cancer signatures on the examined data. Note that the second approach may reflect the prognostic potential of the found pathways, but not the biological convergence of the signatures.

Methods

Compared prognostic signatures

van't Veer's signature was developed based on ~5000 probes (we reproduced a list of 5159 probes) from the Rosetta Hu25K microarray, Wang's signature was developed based on 17816 probes from the Affymetrix U133A microarray. These probes were selected by filtering out probes with low signal, and hence were the actual candidates for the signature, and therefore we chose these lists as the background references.

The Hu25K probes were matched to known genes by their sequences using BLAST [53], and mapped into official gene symbols. We used Affymetrix's mapping of the U133A probe sets to gene symbols. For TANGO analysis probes were converted to Entrez GeneID using MatchMiner [54]. Since not all probes capture a recognized gene with an official gene symbol, and some probes capture more than one gene, the actual lengths of the lists are slightly different than the corresponding list of probes. This, however, does not affect the enrichment analysis as probes with no recognized official gene symbol also have no known annotations. The gene lists of the van't Veer and Wang signatures are listed in Table S1.

Testing for significant pathway enrichment of each list with standard FDR control

The pathway databases used for our analysis are the Gene Ontology Biological Process [45] annotations, (as downloaded from [55]) and MSigSB C2 Canonical Pathways database (version 2.5) [56], which integrates 12 different pathway databases. When referring to a GOBP annotation we refer to all the genes in this annotation together with all the genes of all the descendent annotations. Only annotations that had at least one gene in the relevant background were considered. When considering the size of the annotation, only genes that appear in the relevant background were counted.

For every individual gene list (signature) studied we tested enrichment by genes that belong to a particular biological process or pathway. Enrichment of annotations were computed by Fisher's exact test, using for each signature as background reference the gene population of the original experiment from which the signature was derived (i.e. genes from the corresponding chip that have passed the initial filtering), and correcting for multiple testing by standard control of the FDR [43], without taking the nested dependencies of the GO annotations into account.

More accurate control of multiple hypotheses, using resampling

One might claim that using the standard methods to control the FDR is too strict (mainly due to the dependencies between the pathways). Alternative approaches were suggested to test for annotation enrichments, which were claimed to be less stringent than the standard control of FDR, while still offering correction for testing multiple hypotheses. TANGO [48] performs functional enrichment tests that fully account for multiple testing, using a simple resampling algorithm. The aim is to assess the significance of the enrichment of a gene set T in the different biological processes A_i of an ontology A. First, TANGO computes the hypergeometric p-values p_i of T against all the processes. To determine (in a way that takes multiple testing into account) which of these is significant (at say 5% level), TANGO calculates the empirical background distribution of the best p-values obtained for each one of a large number of randomly generated gene lists (of the same length as T). Finally, the corrected p-value of each process A_i is determined as the probability to do better, using the background distribution. This way all the relations among the biological processes of the ontology are preserved.We have used the EXPANDER [57] implementation of TANGO to test for GO annotation enrichments in all three lists (Wang 60, Wang 76, and van't Veer 70), and determined the threshold on the corrected p-values one needs to use in order to have even a single enriched process shared by van't Veer and one of the Wang lists.

Correcting pathway overlap for multiple testing by assessing the significance of shared processes

At the opposite end of the spectrum of stringency one is ignoring the problem of multiple hypotheses and simply looks for biological processes that passed some threshold on the enrichment p-value *for both gene lists*, such as performing Fisher's exact test [42] and taking only p-values smaller than 0.05. Clearly, since the set of biological processes that satisfied this criterion was derived neglecting completely multiple testing (e.g. of testing many biological processes for enrichment), this procedure is too permissive. To estimate the significance of the fact that a biological process passed this criterion in a way that corrects for multiple testing, we devised a random model to generate a relevant background distribution, which takes into account the real dependences between the pathways and biological processes. Two random lists, *L60* and *L70*, containing 60 and 70 genes, were generated from the respective lists of probes from the chips used by van't Veer and by Wang. We then preformed Fisher's exact test between the genes that correspond to the selected probes of each of the two lists and every biological process (or pathway), and determined the number x of processes with enrichment p-value smaller than some threshold q (we used $q = 0.05$, 0.10 and 1.0), for *both L60* and *L70* (as opposed to *TANGO* which estimates enriched pathways for

every list). By repeating this process 5000 times and calculating the histogram of x, we constructed a background distribution $P(x = k)$, estimating the probability to get by chance k processes with hypergeometric enrichment p-value$<q$ for both random gene lists. Hence, the significance of observing c biological processes or pathways for which both the Wang and van't Veer gene lists are enriched at this level, is simply estimated by $P(x >= c)$. Note, that just like in the actual process of learning the signatures, probes that do not map to known genes with known annotations could be selected, and therefore the effective length of the gene list is usually smaller.

Results

Testing for significant pathway enrichment of each list with standard FDR control

In order to check rigorously the claims of convergent biological pathways and processes for different gene lists, we examined (see Methods) the two most established outcome prediction signatures, the 70 gene list of van't Veer et al. [7] and Wang et al's ER+ signature [33].

We have chosen the richest, well accepted annotation database, the gene ontology biological process (GOBP) database [45] as the major list of pathways, and repeated the analysis with two more lists, for the sake of completeness (see below).

Only a single process, DNA Replication, passed FDR in both signatures, at a very permissive level of 0.31. Raising the bar to FDR = 0.53 gave rise to the microtubule cytoskeleton organization pathway, and with even more permissive FDR only two closely related annotations emerged - 'microtubule-based process' and 'DNA-dependent DNA replication'. This indicates that probably both signatures capture some aspects of cell cycle and proliferation.

It is worth mentioning that the well accepted DAVID annotation tool [58,59] does not find any enriched pathway in any of the signatures, which passed FDR of 0.9, other than organelle organization and biogenesis in Wang's signature (q = 0.36).

We repeated the analysis for a shorter list of GOBPs along the lines of Yu et al. [40] who tested only those 304 GOBPs that had representative probe sets for at least 10 of their genes on the U133A chip. We found 1373 such GOBPs and repeated our analysis limited to this list (we believe that the discrepancy between 304 and 1374 is due to the fact that Yu et al used a very early version of the GO database). Next, we also examined the MSigDB canonical pathways database [56], collecting metabolic and signaling pathways from 12 online pathway databases. Furthermore, we repeated the analysis for the entire signature of Wang (76 probe sets). All these additional comparisons yielded even less common pathways than the original one. The full results of all the pathways that passed FDR of 0.75 in any one of the three databases are shown in Table S2. Few more details can be found in the Methods section.

More accurate control of multiple hypotheses, using resampling

TANGO [48], a resampling based method for pathway enrichment analysis (see Methods), did not find any pathways with p-value smaller than 0.48, see Table S2 for annotations with less significant p-values (DNA Replication was found in all signatures, but with a p-value higher than 0.8).

Correcting pathway overlap for multiple testing by assessing the significance of shared processes. Using the "p-value smaller than 0.05" criteria with p-values obtained by Fisher's exact test of both signatures (see Methods), gave rise to 18 common pathways, most of which were related to cell cycle. Raising the allowed p-value threshold to 0.1 discovers 10 more pathways of different contexts, as also shown in Table 1.

Since this overlap was derived neglecting multiple testing completely, it is too permissive. We estimated the significance of this overlap using a random model (see Methods) to generate a relevant background distribution that takes into account also the real dependences between the pathways and biological processes. This analysis finds that the number of common Biological Processes (derived without any FDR control) that we found for the real lists was significantly higher than the number for random signatures- we calculated a p-value of 0.015 to get the observed overlap for threshold of $p<0.05$ (and 0.068 for $p<0.1$), showing that indeed both signatures capture some common essence. As before, the process was repeated for the reduced GOBP list and MSigDB, as well as for Wang's complete 76 genes signature. The results of the analysis were similar, as shown in Table S3.

What common pathways are really present, other than proliferation?

The fact that both signatures capture cell cycle and proliferation is evident. It is well known that there are many genes whose mRNA level correlates with proliferation, usually referred to as the "proliferation cluster", since they are all cluster together [60,61,62]. Indeed both signatures contain genes from the proliferation cluster, which enables them to approximately capture the rate of proliferation. To test whether there are any additional common pathways, we omitted from both lists the genes that were highly correlated with cell proliferation. Those genes were identified by calculating the Pearson correlation of their expression with the expression of a gene known to be correlated with the rate

Table 1. The list of pathways whose hypergeometric p-value is less than 0.05 and 0.1, without correcting for multiple hypothesis testing.

Common pathways for p<0.05	Additional common pathways for p<0.1
DNA metabolic process	axon regeneration
DNA packaging	cell cycle
DNA replication	cellular component organization
DNA replication initiation	chromatin assembly or disassembly
DNA strand elongation	intracellular signaling cascade
DNA strand elongation during DNA replication	mitotic cell cycle
DNA-dependent DNA replication	negative regulation of translation
cell division	nucleus localization
chromosome condensation	response to hypoxia
chromosome organization	second-messenger-mediated signaling
cytokinesis during cell cycle	
cytoskeleton organization	
microtubule cytoskeleton organization	
microtubule-based process	
mitotic chromosome condensation	
nucleosome assembly	
organelle organization	
phosphoinositide-mediated signaling	

doi:10.1371/journal.pone.0017795.t001

of proliferation in the EMC-344 cohort [33,63]. To be on the safe side, 3 attempts were made, each using a different proliferation gene (either MKI67, TOP2A or CDK1, all of which appear in three major papers discussing the members of the proliferation cluster [60,61,62]). For each attempt, all the genes with Spearman correlation of at least 50% were omitted from the list on which they appeared. The genes that were omitted are listed in Table S4; as can be seen, in all cases the common gene cyclin E2 was omitted. The enrichment analyses described above were now repeated for the filtered signatures.

The results of the enrichment analysis have changed dramatically. No common pathways have passed any FDR (as a matter of fact, no pathway was found in the reduced van't Veer signature that passed any FDR$<$1). Ignoring FDR corrections, only 1–2 common pathways were found with $p<0.05$ (nucleus localization and in the MKI67 case, also response to hypoxia), and 5 common pathways for $p<0.1$. These overlaps were found to be not statistically significant when comparing to the generated background distribution, as described above (p-values of 0.6–0.8). As before, the process was repeated for the reduced GOBP list and MSigDB, as well as for Wang's complete 76 genes signature, yielding similar results. For more details see Table S5 and Table S6.

One might claim that pathway enrichment is not an accurate enough tool to answer the question whether two signatures capture the same biological features. This might be true, but in this case some other proof is necessary, and none has been presented yet. For example, possibly the presence of even one single gene from a pathway could suffice to capture a biological feature, at least to some extent. Proliferation is a good example of such a possibility, since apparently any gene picked from the proliferation cluster will capture the proliferation rate. If this is indeed the case, our test would find an insignificant enrichment, while in fact the pathway *is* represented to some extent. It seems hard to believe, however, that the expression level of one gene can capture the level of activity or deregulation of more complex pathways.

Discussion

We presented a comprehensive analysis aimed at answering the question whether the two outcome prediction signatures for early-discovery breast cancer, of van't Veer et al and Wang et al, capture the same biological processes. We focused on these two signatures since they were derived using machine learning approaches, with minimal biological knowledge incorporated in the choice of the predictive genes. While such an overlap between the biological processes has been claimed or implied, very few studies have actually tested this claim. We performed our tests in a way that on the one hand did not ignore the problems of multiple testing, but on the other hand took into account the dependent and nested nature of the gene ontologies used. We found that the concordance of enriched pathways between the two tested signatures is restricted to capturing the cells' proliferation rate. When proliferation-related genes are deleted from the two lists, the number of pathways over represented in both signatures does not exceed the number of such pathways expected for two random gene lists.

Taken together, all the results obtained indicate that while there is some common biology captured by the two signatures, it is very limited: all the processes captured by both signatures are related to cell proliferation.

To conclude on a constructive note, we do believe that an expression-based prognostic method that is knowledge-based, i.e. one that incorporates also well-established biological and clinical information on relevant pathways, will be able to improve current prediction capabilities.

Supporting Information

Table S1 Gene symbols of the genes in van't Veer and Wang signatures used in the analysis. The Wang signatures were converted from the published probe sets by Affymetrix official tables. The van't Veer signature was converted from the published probes using BLAST. The common gene cyclin E2 is highlighted.
(XLS)

Table S2 Enrichment analysis of each signature separately. FDR controlled hypergeometric enrichment of van't Veer signature, Wang 60 gene ER+ signature and Wang 76 gene signature, for GOBP annotations (both complete and filtered as proposed by Wang et al), and MSigDB pathways. Additionally the results of the TANGO analysis are attached. Pathways common both to van't Veer and one of Wang signatures are highlighted.
(XLS)

Table S3 Pathway overlap significance. The results of our suggested random background model, estimating overlap significance.
(XLS)

Table S4 The proliferation genes omitted. The genes were selected according to correlation in expression in the EMC-344 cohort to the genes MKI67, TOP2A or CDK1.
(XLS)

Table S5 Enrichment analysis after omitting proliferation genes. Hypergeometric enrichment of the signatures minus the genes that correlated with proliferation.
(XLS)

Table S6 Pathway overlap significance after omitting proliferation genes. Same as Table S2, but calculated after omitting proliferation genes.
(XLS)

Acknowledgments

We thank Dr. R. Shen for most helpful correspondence.

Author Contributions

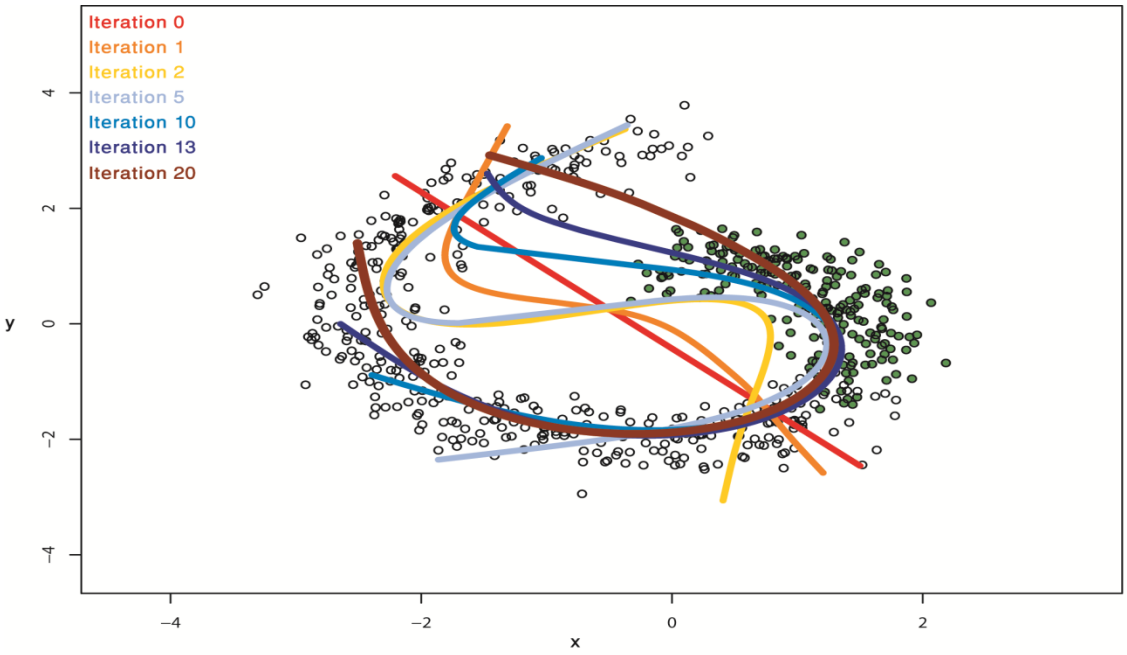
Conceived and designed the experiments: YD ED. Analyzed the data: YD. Wrote the paper: YD ED.

References

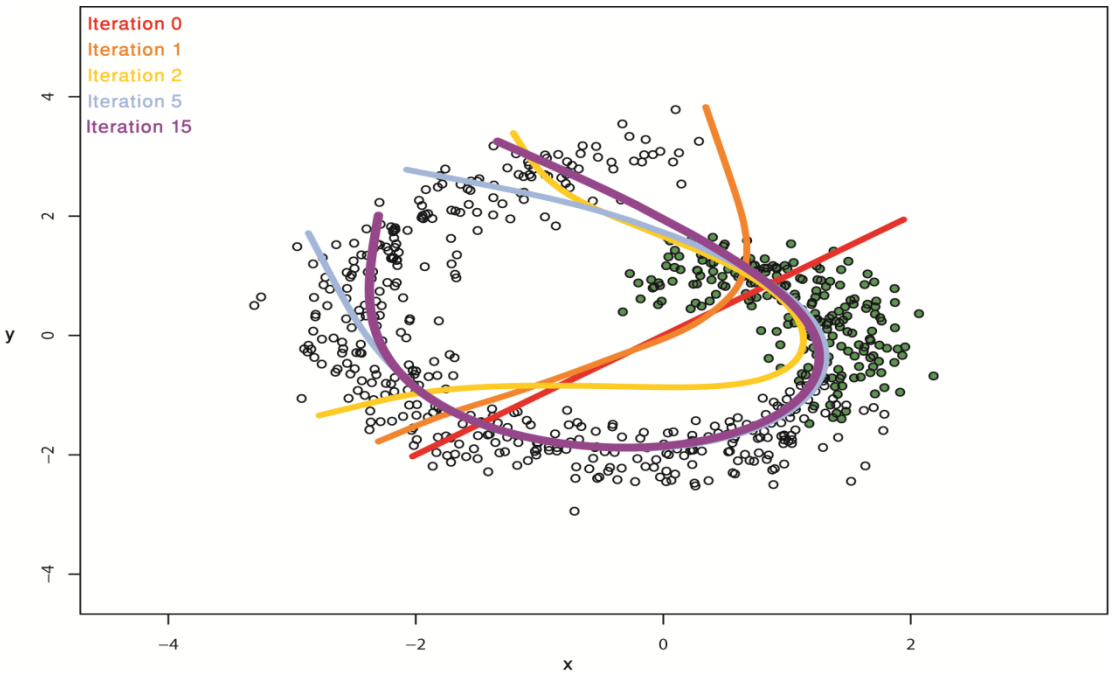
1. van der Vegt B, de Bock GH, Hollema H, Wesseling J (2009) Microarray methods to identify factors determining breast cancer progression: potentials, limitations, and challenges. Crit Rev Oncol Hematol 70: 1–11.
2. Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99: 147–157.
3. Sotiriou C, Pusztai L (2009) Gene-expression signatures in breast cancer. N Engl J Med 360: 790–800.
4. Pusztai L (2009) Gene expression profiling of breast cancer. Breast Cancer Res 11 Suppl 3: S11.
5. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826.
6. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999–2009.

7. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
8. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, et al. (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98: 1183–1192.
9. de Snoo F, Bender R, Glas A, Rutgers E (2009) Gene expression profiling: decoding breast cancer. Surg Oncol 18: 366–378.
10. Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21: 171–178.
11. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365: 488–492.
12. Simon R (2008) Lost in translation: problems and pitfalls in translating laboratory observations to clinical utility. Eur J Cancer 44: 2707–2713.
13. Ransohoff DF (2003) Gene-expression signatures in breast cancer. N Engl J Med 348: 1715–1717; author reply 1715–1717.
14. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928.
15. Kim SY (2009) Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. BMC Bioinformatics 10: 147.
16. Taylor JM, Ankerst DP, Andridge RR (2008) Validation of biomarker-based risk prediction models. Clin Cancer Res 14: 5977–5983.
17. Koscielny S (2008) Critical review of microarray-based prognostic tests and trials in breast cancer. Curr Opin Obstet Gynecol 20: 47–50.
18. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G (2008) A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics 24: 2200–2208.
19. Dunkler D, Michiels S, Schemper M (2007) Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? Eur J Cancer 43: 745–751.
20. Eden P, Ritz C, Rose C, Ferno M, Peterson C (2004) "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Cancer 40: 1837–1841.
21. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, et al. (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355: 560–569.
22. Koscielny S (2010) Why Most Gene Expression Signatures of Tumors Have Not Been Useful in the Clinic. Science Translational Medicine 2: 14ps12.
23. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2: E7.
24. Miller LD, Smeds J, George J, Vega VB, Vergara L, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102: 13550–13555.
25. Takahashi S, Moriya T, Ishida T, Shibata H, Sasano H, et al. (2008) Prediction of breast cancer prognosis by gene expression profile of TP53 status. Cancer Sci 99: 324–332.
26. Troester MA, Herschkowitz JI, Oh DS, He X, Hoadley KA, et al. (2006) Gene expression patterns associated with p53 status in breast cancer. BMC Cancer 6: 276.
27. Glinsky GV, Berezovska O, Glinskii AB (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115: 1503–1521.
28. Paik S, Tang G, Shak S, Kim C, Baker J, et al. (2006) Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 24: 3726–3734.
29. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, et al. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 102: 3738–3743.
30. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98: 262–272.
31. Sotiriou C, Wirapati P, Loi S, Haibe-Kains B, Desmedt C, et al. (2006) Comprehensive analysis integrating both clinicopathological and gene expression data in more than 1,500 samples: Proliferation captured by gene expression grade index appears to be the strongest prognostic factor in breast cancer (BC). J Clin Oncol (Meeting Abstracts) 24: 507-.
32. Sotiriou C, Piccart MJ (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7: 545–553.
33. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679.
34. van't Veer LJ, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452: 564–570.
35. Radpour R, Barekati Z, Kohler C, Holzgreve W, Zhong XY (2009) New trends in molecular biomarker discovery for breast cancer. Genet Test Mol Biomarkers 13: 565–571.
36. Sims AH (2009) Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us? J Clin Pathol 62: 879–885.
37. Reyal F, van Vliet MH, Armstrong NJ, Horlings HM, de Visser KE, et al. (2008) A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. Breast Cancer Res 10: R93.
38. Shen R, Chinnaiyan AM, Ghosh D (2008) Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. BMC Med Genomics 1: 28.
39. Sole X, Bonifaci N, Lopez-Bigas N, Berenguer A, Hernandez P, et al. (2009) Biological convergence of cancer signatures. PLoS One 4: e4544.
40. Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, et al. (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer 7: 182.
41. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, et al. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 8: 37–49.
42. Fisher RA (1970) Statistical methods for research workers. Edinburgh: Oliver & Boyd. xv, 362.
43. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
44. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. Bmj 310: 170.
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
46. Farcomeni A (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res 17: 347–388.
47. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509–515.
48. Tanay A Computational Analysis of Transcriptional Programs: Function and Evolution: PhD Thesis, Tel Aviv University.
49. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
50. Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, et al. (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Med 3: e47.
51. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C (2007) An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 8: R157.
52. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. Nature 406: 747–752.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
54. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, et al. (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol 4: R27.
55. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, et al. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Res 37: D396–403.
56. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
57. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, et al. (2005) EXPANDER–an integrative program suite for microarray data analysis. BMC Bioinformatics 6: 232.
58. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.
59. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
60. Rosty C, Sheffer M, Tsafrir D, Stransky N, Tsafrir I, et al. (2005) Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. Oncogene 24: 7094–7104.
61. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24: 227–235.
62. Whitfield ML, George LK, Grant GD, Perou CM (2006) Common markers of proliferation. Nat Rev Cancer 6: 99–106.
63. Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, et al. (2007) Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci U S A 104: 6740–6745.

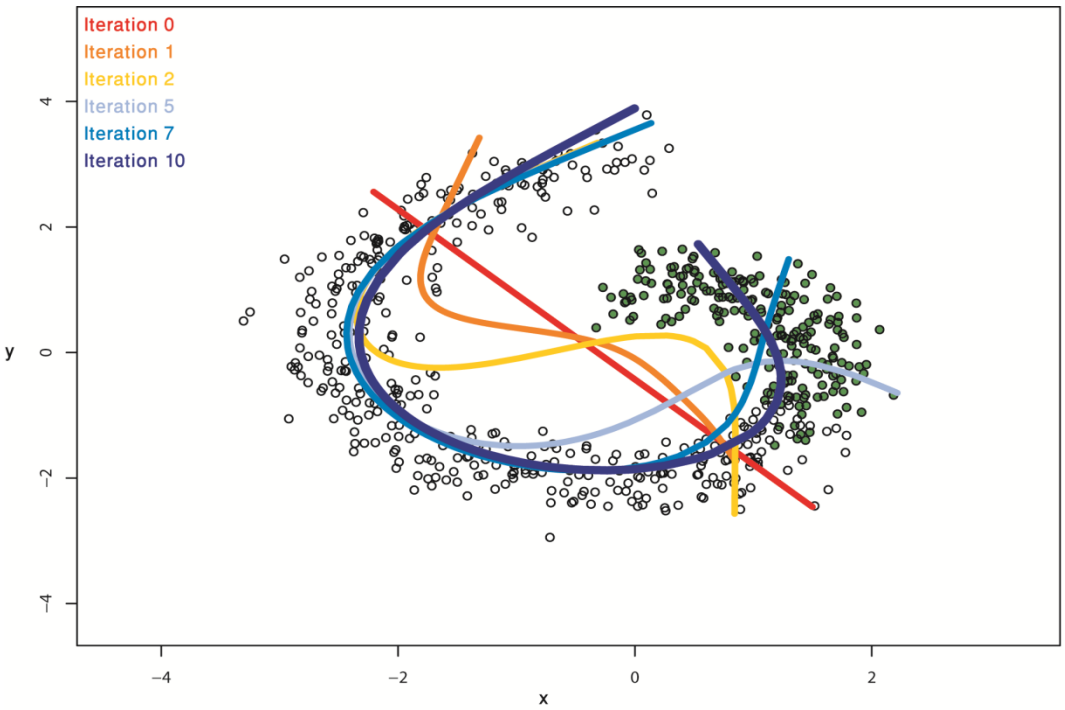
Appendix B – Chapter 6 Supplementary Figures And Tables



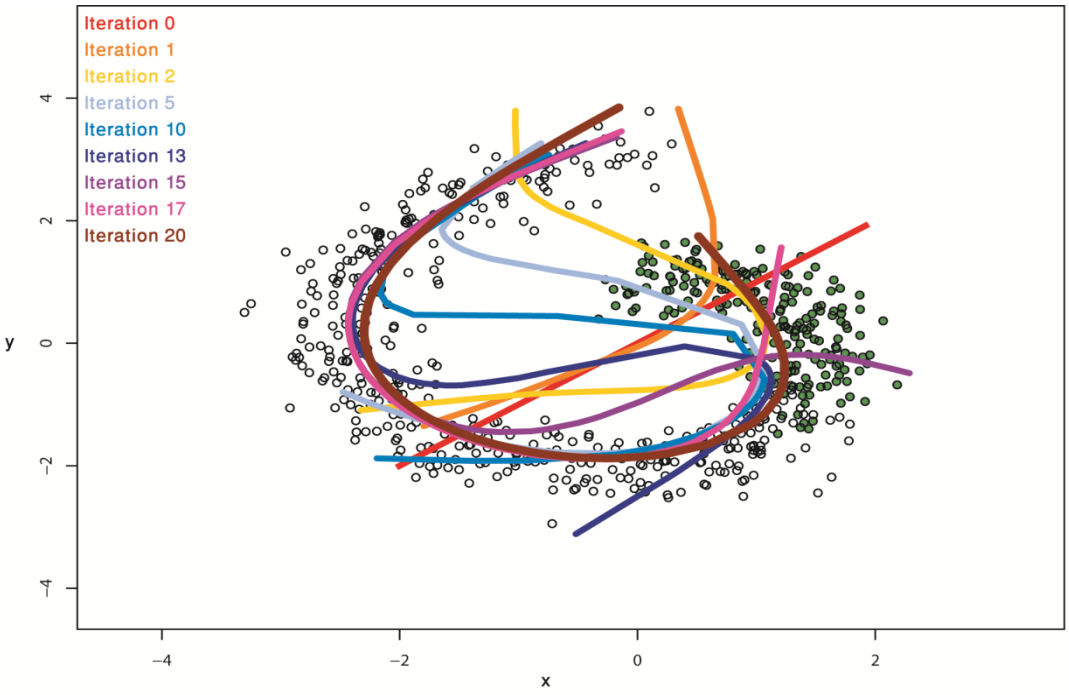
Supplementary Figure 1 - **Hastie and Stuetzle's principal curve for simulated noisy trajectory**. After 20 iterations the algorithm converges to the (wrong) curve shown in brown. The initial guess (first principal component) is shown in red. *t=0* points are in green.



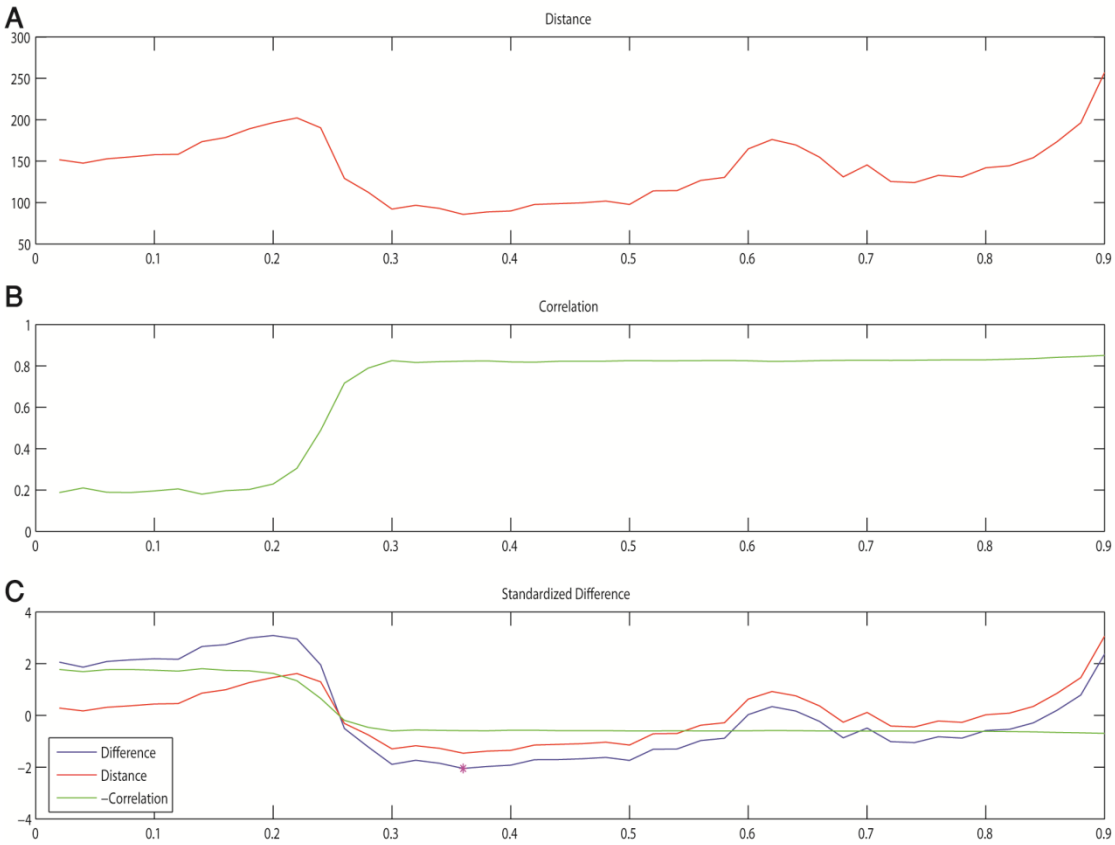
Supplementary Figure 2 - **Hastie and Stuetzle's principal curve for simulated noisy trajectory, with an educated initial guess**. The initial starting curve (in red) reflects external knowledge and goes through the centers of the points labeled as *t=0* (in green) and *t=1*. After 15 iterations the algorithm converges to the (wrong) purple curve.



Supplementary Figure 3 – **Modified semi-supervised principal curve for simulated noisy trajectory**. After 10 iterations the algorithm converges to the curve shown in dark blue, capturing the real trajectory. The initial guess (first principal component) is shown in red. *t=0* points are shown in green.



Supplementary Figure 4 – **Modified semi-supervised principal curve for simulated noisy trajectory**. After 20 iterations the algorithm converges to the curve shown in brown, capturing the real trajectory. The initial guess (line that goes through the center of *t=0* and *t=1* points) is shown in red. *t=0* points are shown in green.



Supplementary Figure 5 – **Finding an optimal α for the simulated data. A.** The total distance to the curve as a function of α. **B.** The correlation between the curve parameter and the given ranking. **C.** The two standardized terms (correlation is negated for easier view) and their difference. For any $\alpha \geq 0.28$ the correct curve is captured and therefore the correlation is high, while for $\alpha \leq 0.2$ a curve similar to the Hastie and Stuetzle's principal curve is captured. The minimum is achieved for $\alpha = 0.36$ (marked by pink star). The correlation term makes sure that $\alpha \geq 0.28$ is selected and the distance term ensures that the selected α will not be too high and disturb the curve too much.

Pathway Number	Pathway name	Gene Mutated	p value	FDR	Cluster
1	pid alk2pathway	IDH1	7.50E-05	6.32E-03	
2	biocarta hsp27 pathway	IDH1	8.50E-05	6.47E-03	1
3	kegg peroxisome	IDH1	8.20E-05	6.47E-03	1
4	kegg abc transporters	IDH1	1.00E-06	1.24E-03	1
5	pid alphasynuclein pathway	IDH1	4.60E-05	4.72E-03	1
6	kegg retinol metabolism	DST	1.32E-04	8.01E-03	1
7	kegg histidine metabolism	IDH1	1.00E-06	1.24E-03	1
8	pid cmyb pathway	RB1	2.90E-05	3.69E-03	
9	pid ap1 pathway	RB1	1.00E-05	2.03E-03	
9	pid ap1 pathway	TP53	8.90E-05	6.53E-03	
10	biocarta tgfb pathway	IDH1	1.14E-04	7.19E-03	
11	kegg calcium signaling pathway	EGFR	8.60E-05	6.47E-03	2
12	biocarta spry pathway	IDH1	7.50E-05	6.32E-03	2
13	biocarta keratinocyte pathway	EGFR	4.00E-06	1.96E-03	2
14	pid upa upar pathway	EGFR	6.00E-06	1.96E-03	2
15	pid lysophospholipid pathway	EGFR	9.80E-05	6.94E-03	2
16	biocarta at1r pathway	EGFR	1.05E-04	7.11E-03	2
17	biocarta cbl pathway	EGFR	7.00E-06	1.96E-03	2
18	pid arf6 pathway	EGFR	1.08E-04	7.11E-03	2
19	biocarta cardiacegf pathway	EGFR	5.00E-06	1.96E-03	2
20	pid erbb1 receptor proximal pathway	EGFR	2.50E-05	3.45E-03	2
21	pid syndecan 3 pathway	EGFR	2.40E-05	3.45E-03	2
22	biocarta tff pathway	EGFR	2.60E-05	3.53E-03	2
23	biocarta her2 pathway	EGFR	2.10E-05	3.33E-03	2
24	pid erbb network pathway	EGFR	7.70E-05	6.32E-03	2
25	biocarta mcalpain pathway	EGFR	1.80E-05	3.00E-03	2
26	pid telomerasepathway	EGFR	7.00E-06	1.96E-03	2
27	kegg endometrial cancer	EGFR	2.50E-05	3.45E-03	2
28	pid ecadherin keratinocyte pathway	EGFR	8.60E-05	6.47E-03	2
29	biocarta erk pathway	EGFR	1.30E-05	2.32E-03	2
30	pid txa2pathway	EGFR	8.00E-06	1.96E-03	2
31	biocarta egf pathway	EGFR	1.00E-05	2.03E-03	2
32	kegg dorso ventral axis formation	EGFR	8.00E-05	6.47E-03	2
33	pid a6b1 a6b4 integrin pathway	EGFR	4.00E-06	1.96E-03	2
33	pid a6b1 a6b4 integrin pathway	IDH1	4.30E-05	4.66E-03	2
34	biocarta egfr smrte pathway	EGFR	3.00E-06	1.96E-03	2
34	biocarta egfr smrte pathway	IDH1	3.90E-05	4.55E-03	2
35	pid ptp1bpathway	EGFR	4.40E-05	4.67E-03	2
36	pid ecadherin stabilization pathway	EGFR	2.00E-06	1.91E-03	2
37	kegg epithelial cell signaling in helicobacter pylori infection	EGFR	6.00E-06	1.96E-03	2
37	kegg epithelial cell signaling in helicobacter pylori infection	IDH1	1.72E-04	9.59E-03	2
38	pid erbb1 downstream pathway	EGFR	2.80E-05	3.61E-03	2
38	pid erbb1 downstream pathway	IDH1	1.47E-04	8.66E-03	2

39	biocarta agr pathway	EGFR	1.70E-05	3.00E-03	2
40	kegg erbb signaling pathway	EGFR	2.00E-06	1.91E-03	2
41	kegg gap junction	EGFR	2.50E-05	3.45E-03	2
42	kegg mapk signaling pathway	IDH1	7.70E-05	6.32E-03	2
43	pid retinoic acid pathway	IDH1	1.00E-06	1.24E-03	
44	biocarta pitx2 pathway	IDH1	1.00E-06	1.24E-03	
45	biocarta alk pathway	IDH1	7.00E-06	1.96E-03	
46	biocarta igf1 pathway	NF1	1.21E-04	7.53E-03	3
47	pid erbb2erbb3pathway	NF1	1.59E-04	9.07E-03	3
48	biocarta tcapoptosis pathway	IDH1	1.56E-04	9.02E-03	3
49	kegg systemic lupus erythematosus	RB1	5.00E-06	1.96E-03	3
50	pid pdgfrapathway	NF1	8.90E-05	6.53E-03	3
51	pid ceramide pathway	IDH1	4.00E-05	4.55E-03	3
52	biocarta sodd pathway	IDH1	1.24E-04	7.61E-03	3
53	pid syndecan pathway	IDH1	4.00E-05	4.55E-03	3
54	kegg propanoate metabolism	IDH1	9.00E-06	2.03E-03	3
55	kegg long term depression	IDH1	4.10E-05	4.60E-03	3
56	kegg alpha linolenic acid metabolism	IDH1	5.30E-05	5.16E-03	3
57	kegg ether lipid metabolism	IDH1	1.90E-05	3.22E-03	3
58	kegg glycolysis gluconeogenesis	IDH1	3.60E-05	4.49E-03	3
59	pid ret pathway	IDH1	1.33E-04	8.04E-03	3
60	pid integrin4 pathway	IDH1	1.52E-04	8.85E-03	3
61	pid hedgehog glipathway	IDH1	7.20E-05	6.30E-03	3
62	biocarta p38mapk pathway	IDH1	5.70E-05	5.22E-03	3
63	pid integrin cs pathway	NF1	1.79E-04	9.83E-03	3
64	pid tap63pathway	IDH1	1.14E-04	7.19E-03	3
64	pid tap63pathway	NF1	4.30E-05	4.66E-03	3
65	pid reelinpathway	NF1	1.30E-05	2.32E-03	3
66	pid cxcr4 pathway	IDH1	1.77E-04	9.81E-03	3
67	kegg hypertrophic cardiomyopathy hcm	NF1	6.60E-05	5.80E-03	3
68	biocarta extrinsic pathway	NF1	6.30E-05	5.67E-03	3
69	pid wnt signaling pathway	NF1	2.30E-05	3.45E-03	3
70	pid ps1pathway	NF1	1.09E-04	7.12E-03	3
71	kegg wnt signaling pathway	NF1	5.60E-05	5.21E-03	3
72	pid il4 2pathway	PTEN	1.13E-04	7.19E-03	3
73	pid wnt noncanonical pathway	IDH1	1.38E-04	8.22E-03	3
74	kegg focal adhesion	IDH1	1.02E-04	7.06E-03	3
75	pid nfkappabcanonicalpathway	IDH1	8.30E-05	6.47E-03	3
76	pid syndecan 4 pathway	IDH1	5.00E-05	5.03E-03	3
77	kegg leukocyte transendothelial migration	IDH1	4.00E-05	4.55E-03	3
78	kegg cell adhesion molecules cams	IDH1	1.07E-04	7.11E-03	3
79	kegg chemokine signaling pathway	IDH1	2.00E-05	3.22E-03	3
80	kegg cytokine cytokine receptor interaction	IDH1	5.40E-05	5.16E-03	3
81	kegg glycosaminoglycan degradation	NF1	1.04E-04	7.11E-03	3

82	kegg other glycan degradation	IDH1	8.00E-06	1.96E-03	3
83	pid igf1 pathway	IDH1	1.83E-04	9.93E-03	3
83	pid igf1 pathway	NF1	1.67E-04	9.43E-03	3
84	biocarta longevity pathway	IDH1	6.00E-06	1.96E-03	3
85	pid angiopoietinreceptor pathway	NF1	4.70E-05	4.80E-03	3
86	pid il3 pathway	NF1	5.50E-05	5.16E-03	3
87	biocarta igf1r pathway	NF1	1.00E-05	2.03E-03	3
88	biocarta pyk2 pathway	NF1	1.00E-05	2.03E-03	3
89	biocarta insulin pathway	NF1	8.00E-06	1.96E-03	3
90	biocarta il3 pathway	NF1	8.00E-06	1.96E-03	3
91	biocarta ngf pathway	NF1	7.00E-06	1.96E-03	3
92	biocarta epo pathway	NF1	7.00E-06	1.96E-03	3
93	biocarta trka pathway	NF1	9.90E-05	6.94E-03	3
94	biocarta erk5 pathway	NF1	9.40E-05	6.77E-03	3

Supplementary Table 1 - **Pathways whose deregulation corresponds to point mutation of selected genes (TCGA GBM data).** Pathways are ordered and numbered as in Figure 6-2-A.

Pathway	p-value	FDR	Correlation Coefficient
pid hif2pathway	1.54E-19	8.43E-17	0.41
pid hif1 tfpathway	1.60E-18	4.39E-16	0.40
pid integrin a9b1 pathway	1.45E-17	2.64E-15	0.39
pid s1p s1p1 pathway	5.14E-17	7.04E-15	0.38
kegg nitrogen metabolism	3.21E-16	3.52E-14	0.38
biocarta hif pathway	2.64E-15	2.41E-13	0.36
pid vegf vegfr pathway	9.34E-15	7.28E-13	0.36
pid fra pathway	1.06E-14	7.28E-13	0.36
biocarta no1 pathway	4.93E-14	3.00E-12	0.35
pid vegfr1 2 pathway	8.91E-14	4.89E-12	0.34
pid vegfr1 pathway	1.04E-13	5.17E-12	0.34
kegg mtor signaling pathway	5.52E-13	2.52E-11	0.33
pid rxr vdr pathway	7.07E-13	2.98E-11	0.33
kegg propanoate metabolism	9.96E-13	3.90E-11	0.33
pid integrin3 pathway	1.61E-12	5.90E-11	0.33
biocarta vegf pathway	3.19E-12	1.09E-10	0.32
pid il1pathway	5.99E-12	1.93E-10	0.32
pid endothelinpathway	1.34E-11	4.09E-10	0.31
kegg ppar signaling pathway	7.02E-11	2.02E-09	0.30
pid cd40 pathway	1.49E-10	3.99E-09	0.30
kegg valine leucine and isoleucine degradation	1.53E-10	3.99E-09	0.30
biocarta mitochondria pathway	2.13E-10	5.21E-09	0.30
kegg apoptosis	2.19E-10	5.21E-09	0.30
biocarta sodd pathway	3.04E-10	6.94E-09	0.29
pid syndecan 4 pathway	7.65E-10	1.61E-08	0.29
pid faspathway	7.77E-10	1.61E-08	0.29
pid hivnefpathway	7.94E-10	1.61E-08	0.29
pid ceramide pathway	8.23E-10	1.61E-08	0.29
pid p38alphabetadownstreampathway	1.28E-09	2.42E-08	0.28
kegg melanogenesis	1.33E-09	2.43E-08	0.28
kegg beta alanine metabolism	1.90E-09	3.36E-08	0.28
kegg alanine aspartate and glutamate metabolism	2.15E-09	3.68E-08	0.28
pid p75ntrpathway	4.21E-09	6.99E-08	0.27
kegg neuroactive ligand receptor interaction	4.70E-09	7.57E-08	0.27
kegg wnt signaling pathway	7.70E-09	1.18E-07	0.27
kegg neurotrophin signaling pathway	7.75E-09	1.18E-07	0.27
kegg vegf signaling pathway	8.31E-09	1.23E-07	0.27
pid lymphangiogenesis pathway	1.26E-08	1.82E-07	0.27
kegg adipocytokine signaling pathway	1.71E-08	2.37E-07	0.26
kegg nod like receptor signaling pathway	1.73E-08	2.37E-07	0.26
kegg long term potentiation	2.03E-08	2.71E-07	0.26
kegg phosphatidylinositol signaling system	2.31E-08	2.97E-07	0.26
pid caspase pathway	2.36E-08	2.97E-07	0.26
kegg renal cell carcinoma	2.38E-08	2.97E-07	0.26

pid avb3 opn pathway	2.63E-08	3.20E-07	0.26
pid glypican 1pathway	2.83E-08	3.37E-07	0.26
pid avb3 integrin pathway	2.97E-08	3.47E-07	0.26
pid met pathway	3.86E-08	4.41E-07	0.26
pid tcrcalciumpathway	5.60E-08	6.14E-07	0.25
pid angiopoietinreceptor pathway	5.60E-08	6.14E-07	0.25
pid foxopathway	6.16E-08	6.56E-07	0.25
pid integrin a4b1 pathway	6.23E-08	6.56E-07	0.25
pid syndecan 1 pathway	8.39E-08	8.64E-07	0.25
pid fgf pathway	8.52E-08	8.64E-07	0.25
pid integrin1 pathway	9.10E-08	9.06E-07	0.25
kegg inositol phosphate metabolism	9.34E-08	9.14E-07	0.25
kegg glycolysis gluconeogenesis	9.94E-08	9.55E-07	0.25
pid p53downstreampathway	1.16E-07	1.09E-06	0.25
kegg glycosphingolipid biosynthesis globo series	1.18E-07	1.10E-06	0.25
pid notch pathway	1.59E-07	1.43E-06	0.25
kegg fatty acid metabolism	2.35E-07	2.08E-06	0.24
kegg glycosphingolipid biosynthesis lacto and neolacto series	2.39E-07	2.08E-06	0.24
biocarta p38mapk pathway	2.56E-07	2.18E-06	0.24
kegg cytokine cytokine receptor interaction	2.58E-07	2.18E-06	0.24
pid kitpathway	2.65E-07	2.20E-06	0.24
biocarta fibrinolysis pathway	4.07E-07	3.33E-06	0.24
pid il2 pi3kpathway	6.33E-07	5.10E-06	0.23
pid botulinumtoxinpathway	7.00E-07	5.56E-06	0.23
biocarta gaba pathway	8.11E-07	6.35E-06	0.23
kegg butanoate metabolism	8.55E-07	6.60E-06	0.23
kegg lysine degradation	8.85E-07	6.74E-06	0.23
biocarta cacam pathway	1.02E-06	7.67E-06	0.23
kegg insulin signaling pathway	1.04E-06	7.67E-06	0.23
pid fak pathway	1.08E-06	7.78E-06	0.23
pid pi3kplctrkpathway	1.08E-06	7.78E-06	0.23
kegg chemokine signaling pathway	1.18E-06	8.40E-06	0.23
kegg long term depression	1.24E-06	8.68E-06	0.23
pid ephbfwdpathway	1.60E-06	1.11E-05	0.23
pid tgfbrpathway	1.80E-06	1.23E-05	0.22
kegg snare interactions in vesicular transport	2.11E-06	1.43E-05	0.22
pid insulin pathway	3.09E-06	2.04E-05	0.22
biocarta pgc1a pathway	3.13E-06	2.04E-05	0.22
kegg cardiac muscle contraction	3.25E-06	2.09E-05	0.22
biocarta actiny pathway	3.34E-06	2.13E-05	0.22
kegg starch and sucrose metabolism	3.69E-06	2.33E-05	0.22
pid myc represspathway	3.95E-06	2.46E-05	0.22
pid integrin5 pathway	4.10E-06	2.52E-05	0.22
pid il8cxcr2 pathway	4.29E-06	2.61E-05	0.22
pid amb2 neutrophils pathway	4.37E-06	2.61E-05	0.22

pid nfkappabcanonicalpathway	4.40E-06	2.61E-05	0.22
kegg taurine and hypotaurine metabolism	4.43E-06	2.61E-05	0.22
biocarta eif4 pathway	4.57E-06	2.66E-05	0.22
pid syndecan 2 pathway	4.95E-06	2.86E-05	0.22
kegg proximal tubule bicarbonate reclamation	5.24E-06	2.99E-05	0.21
kegg nicotinate and nicotinamide metabolism	5.58E-06	3.15E-05	0.21
kegg ascorbate and aldarate metabolism	6.12E-06	3.42E-05	0.21
biocarta creb pathway	6.48E-06	3.59E-05	0.21
pid trkrpathway	7.50E-06	4.11E-05	0.21
kegg fructose and mannose metabolism	7.83E-06	4.25E-05	0.21
kegg arachidonic acid metabolism	7.93E-06	4.26E-05	0.21
kegg focal adhesion	8.32E-06	4.43E-05	0.21
kegg glycerophospholipid metabolism	8.56E-06	4.51E-05	0.21
pid rhoa pathway	8.70E-06	4.54E-05	0.21
kegg ecm receptor interaction	9.32E-06	4.82E-05	0.21
pid ps1pathway	9.76E-06	5.00E-05	0.21
biocarta arenrf2 pathway	1.00E-05	5.08E-05	0.21
kegg linoleic acid metabolism	1.09E-05	5.49E-05	0.21
biocarta bad pathway	1.16E-05	5.76E-05	0.21
pid wnt noncanonical pathway	1.24E-05	6.12E-05	0.21
pid bcr 5pathway	1.32E-05	6.47E-05	0.21
biocarta caspase pathway	1.60E-05	7.74E-05	0.20
kegg taste transduction	1.69E-05	8.14E-05	0.20
pid cd8tcrdownstreampathway	1.97E-05	9.36E-05	0.20
pid pdgfrbpathway	1.98E-05	9.36E-05	0.20
pid mapktrkpathway	2.14E-05	9.94E-05	0.20
kegg ether lipid metabolism	2.15E-05	9.94E-05	0.20
biocarta ck1 pathway	2.16E-05	9.94E-05	0.20
kegg vascular smooth muscle contraction	2.18E-05	9.94E-05	0.20
pid wnt canonical pathway	3.12E-05	1.41E-04	0.20
kegg glutathione metabolism	3.18E-05	1.43E-04	0.20
biocarta nthi pathway	3.36E-05	1.50E-04	0.20
biocarta insulin pathway	3.43E-05	1.52E-04	0.20
biocarta cftr pathway	3.62E-05	1.59E-04	0.20
pid cxcr3pathway	3.69E-05	1.60E-04	0.19
pid wnt signaling pathway	3.74E-05	1.61E-04	0.19
biocarta plce pathway	3.81E-05	1.63E-04	0.19
pid il8cxcr1 pathway	3.86E-05	1.64E-04	0.19
pid ifngpathway	4.21E-05	1.77E-04	0.19
pid tap63pathway	4.33E-05	1.80E-04	0.19
biocarta cdk5 pathway	4.34E-05	1.80E-04	0.19
pid integrin2 pathway	4.48E-05	1.85E-04	0.19
kegg hematopoietic cell lineage	4.68E-05	1.91E-04	0.19
kegg olfactory transduction	4.71E-05	1.91E-04	0.19
biocarta gh pathway	5.46E-05	2.20E-04	0.19
biocarta ngf pathway	5.93E-05	2.37E-04	0.19

kegg tight junction	6.50E-05	2.58E-04	0.19
kegg glycosaminoglycan biosynthesis heparan sulfate	6.67E-05	2.63E-04	0.19
pid cdc42 pathway	7.12E-05	2.79E-04	0.19
kegg leukocyte transendothelial migration	8.09E-05	3.14E-04	0.19
biocarta tob1 pathway	8.39E-05	3.24E-04	0.19
biocarta th1th2 pathway	8.87E-05	3.40E-04	0.19
kegg arrhythmogenic right ventricular cardiomyopathy arvc	9.02E-05	3.43E-04	0.19
pid arf6 traffickingpathway	9.20E-05	3.48E-04	0.18
biocarta il2rb pathway	9.39E-05	3.52E-04	0.18
biocarta chrebp2 pathway	9.95E-05	3.71E-04	0.18
biocarta biopeptides pathway	1.04E-04	3.85E-04	0.18
pid alk1pathway	1.08E-04	3.96E-04	0.18
kegg alpha linolenic acid metabolism	1.13E-04	4.13E-04	0.18
biocarta integrin pathway	1.17E-04	4.25E-04	0.18
kegg regulation of actin cytoskeleton	1.37E-04	4.95E-04	0.18
biocarta erk5 pathway	1.39E-04	4.95E-04	0.18
biocarta nuclearrs pathway	1.39E-04	4.95E-04	0.18
biocarta epo pathway	1.46E-04	5.17E-04	0.18
biocarta extrinsic pathway	1.47E-04	5.18E-04	0.18
biocarta calcineurin pathway	1.75E-04	6.06E-04	0.18
kegg pathways in cancer	1.77E-04	6.10E-04	0.18
pid cxcr4 pathway	2.00E-04	6.84E-04	0.18
biocarta monocyte pathway	2.01E-04	6.84E-04	0.18
kegg aldosterone regulated sodium reabsorption	2.06E-04	6.96E-04	0.18
biocarta trka pathway	2.08E-04	6.98E-04	0.18
pid syndecan pathway	2.27E-04	7.57E-04	0.17
biocarta il3 pathway	2.42E-04	8.04E-04	0.17
biocarta il7 pathway	2.45E-04	8.08E-04	0.17
kegg gnrh signaling pathway	2.49E-04	8.15E-04	0.17
biocarta ndkdynamin pathway	2.51E-04	8.15E-04	0.17
biocarta tcr pathway	2.51E-04	8.15E-04	0.17
pid a4b7 pathway	2.89E-04	9.33E-04	0.17
biocarta igf1r pathway	2.98E-04	9.55E-04	0.17
biocarta il6 pathway	3.07E-04	9.71E-04	0.17
biocarta barrestin src pathway	3.15E-04	9.92E-04	0.17
biocarta hivnef pathway	3.25E-04	1.02E-03	0.17
biocarta cytokine pathway	3.87E-04	1.20E-03	0.17
kegg glycerolipid metabolism	3.92E-04	1.21E-03	0.17
pid lpa4 pathway	4.03E-04	1.22E-03	0.17
biocarta barr mapk pathway	4.04E-04	1.22E-03	0.17
biocarta barrestin pathway	4.04E-04	1.22E-03	0.17
pid tcrjnkpathway	4.53E-04	1.36E-03	0.17
kegg oxidative phosphorylation	4.97E-04	1.49E-03	0.16
biocarta stathmin pathway	5.03E-04	1.50E-03	0.16
biocarta par1 pathway	5.42E-04	1.60E-03	0.16
biocarta mapk pathway	5.53E-04	1.63E-03	0.16

biocarta longevity pathway	5.93E-04	1.74E-03	0.16
kegg cell adhesion molecules cams	5.96E-04	1.74E-03	0.16
kegg o glycan biosynthesis	6.47E-04	1.87E-03	0.16
biocarta pyk2 pathway	7.33E-04	2.12E-03	0.16
biocarta lair pathway	7.58E-04	2.17E-03	0.16
biocarta il1r pathway	8.92E-04	2.55E-03	0.16
biocarta nkt pathway	9.60E-04	2.72E-03	0.16
kegg n glycan biosynthesis	9.75E-04	2.75E-03	0.16
pid ncadherinpathway	9.96E-04	2.80E-03	0.16
kegg endocytosis	1.02E-03	2.84E-03	0.16
biocarta ace2 pathway	1.04E-03	2.88E-03	0.16
pid ret pathway	1.23E-03	3.40E-03	0.15
pid cd8tcrpathway	1.25E-03	3.45E-03	0.15
pid p38alphabetapathway	1.28E-03	3.52E-03	0.15
biocarta rankl pathway	1.30E-03	3.55E-03	0.15
kegg dilated cardiomyopathy	1.35E-03	3.67E-03	0.15
kegg vasopressin regulated water reabsorption	1.39E-03	3.74E-03	0.15
pid hedgehog glipathway	1.40E-03	3.75E-03	0.15
biocarta p53hypoxia pathway	1.49E-03	3.96E-03	0.15
kegg other glycan degradation	1.49E-03	3.96E-03	0.15
pid p38 mkk3 6pathway	1.58E-03	4.19E-03	0.15
kegg hypertrophic cardiomyopathy hcm	1.60E-03	4.21E-03	0.15
biocarta hdac pathway	1.63E-03	4.27E-03	0.15
biocarta inflam pathway	1.65E-03	4.31E-03	0.15
biocarta vip pathway	1.69E-03	4.39E-03	0.15
pid integrin4 pathway	1.73E-03	4.46E-03	0.15
pid mtor 4pathway	1.84E-03	4.73E-03	0.15
biocarta lym pathway	1.94E-03	4.95E-03	0.15
kegg lysosome	1.94E-03	4.95E-03	0.15
pid cdc42 reg pathway	2.16E-03	5.48E-03	0.15
biocarta akapcentrosome pathway	2.18E-03	5.49E-03	0.15
biocarta ccr5 pathway	2.18E-03	5.49E-03	0.15
biocarta vitcb pathway	2.20E-03	5.51E-03	0.15
kegg ribosome	2.26E-03	5.63E-03	0.14
kegg glycosaminoglycan biosynthesis chondroitin sulfate	2.35E-03	5.83E-03	0.14
biocarta igf1 pathway	2.45E-03	6.02E-03	0.14
kegg huntingtons disease	2.58E-03	6.30E-03	0.14
pid erbb4 pathway	2.60E-03	6.30E-03	0.14
kegg riboflavin metabolism	2.65E-03	6.41E-03	0.14
pid epha2 fwdpathway	2.67E-03	6.41E-03	0.14
biocarta tcapoptosis pathway	2.73E-03	6.53E-03	0.14
pid ilk pathway	2.87E-03	6.84E-03	0.14
pid il3 pathway	2.89E-03	6.84E-03	0.14
kegg amyotrophic lateral sclerosis als	2.89E-03	6.84E-03	0.14
biocarta eponfkb pathway	2.98E-03	7.00E-03	0.14
kegg glycosaminoglycan degradation	2.99E-03	7.00E-03	0.14

pid reg gr pathway	3.10E-03	7.22E-03	0.14
kegg amino sugar and nucleotide sugar metabolism	3.13E-03	7.28E-03	0.14
pid hnf3apathway	3.48E-03	8.01E-03	0.14
kegg complement and coagulation cascades	3.62E-03	8.30E-03	0.14
pid reelinpathway	3.96E-03	9.05E-03	0.14
pid ephrinbrevpathway	4.36E-03	9.90E-03	0.14
pid gmcsf pathway	4.41E-03	9.99E-03	0.14
biocarta met pathway	3.18E-03	7.36E-03	-0.14
pid bmppathway	2.48E-03	6.07E-03	-0.14
kegg rig i like receptor signaling pathway	2.37E-03	5.86E-03	-0.14
kegg glycine serine and threonine metabolism	3.46E-04	1.08E-03	-0.17
biocarta gleevec pathway	3.04E-04	9.69E-04	-0.17
biocarta ps1 pathway	1.73E-04	6.04E-04	-0.18
kegg notch signaling pathway	2.96E-06	1.98E-05	-0.22
biocarta death pathway	1.43E-07	1.31E-06	-0.25

Supplementary Table 2 - **Pathways whose deregulation correlates with necrosis levels** (TCGA GBM data).