Spring 5-1-2013

Identification of Genetic Determinants of Metastasis and Clonal Relationships Between Primary and Metastatic Tumors

Gaurav Singhal
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology Program

Identification of Genetic Determinants of Metastasis and Clonal Relationships Between Primary and Metastatic Tumors

by

Gaurav Singhal

A thesis presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the
degree of Master of Arts

May 2013

St. Louis, Missouri

Acknowledgements

I would like to thank Rakesh Nagarajan for his mentorship and support along with committee members Gary Stormo, Mark Watson and Christopher Maher. Members of the Nagarajan Lab and Genome Technology Access Center, especially Paul Cliften, provided important feedback for experimental design and analysis. Very special thanks go to Mark Watson for providing data and advice for analysis throughout this project. I would also like to thank Barak Cohen for his advice and mentoring as the Program Director of Computational & Systems Biology Program and the Division of Biology and Biomedical Sciences for their training and financial support. Lastly, I would like to thank my parents Ajay Gupta and Anjana Gupta for their patience, encouragement, and unwavering support.

Table of Contents

List of Figures

Background and Significance

Metastasis accounts for greater than 90% of cancer related mortality in solid tissue cancers. Though cancer related mortality has decreased steadily over the past decade, much of the reduction in cancer incidence and mortality has been accounted to come from increased prevalence of preventive care (such as mammography screening) and reduced exposure to risk factors (such as tobacco consumption) [1]. It is widely accepted that the most effective improvements in therapy outcome will come from management of metastasis, or reduction in metastasis related mortality.

Cancer metastasis is widely viewed as an evolutionary process [2], and is believed to have a strong genetic component. Several studies (14-18) have appeared in the past decade which strongly suggest that genetic adaptations play a role in enhancing the metastatic potential of tumor cells. Gene sets have been identified in xenografts of cancer cell line sub-populations selected specifically for elevated metastatic ability (14, 15). Analyses of DNA sequence data from human pancreatic primary cancer and metastatic tissues have indicated that differential potency for initiating metastasis can be partly attributed to the genetic heterogeneity among subpopulations of cells within primary tumors (16).

Multiple studies to date have shown that targeting the effect of a mutation leads to recession of metastasis. For example, recurrent inactivating somatic mutations were identified in BAP1 in highly metastatic uveal melanomas of the eye [17]. RNA interference mediated knock down of

BAP1 in uveal melanoma cells lead to transformation into a de-differentiated class 2 UM phenotype. Also, it has been shown that targeting metastasis-related genes dramatically improves survival and therapy outcome [18, 25].

Another recent study [3] showed that genetically identical tumor cells can display functional variability especially in proliferation persistence and chemotherapy tolerance, suggesting that non-genetic components may also contribute significantly towards cancer evolution and persistence.

Multiple studies have tried to identify genetic determinants of Non Small Cell Lung Cancer [4,5], leading to discovery of recurrently mutated genes and significantly altered pathways. However, these studies have been limited to analysis of primary tumors only, and hardly any effort has been made towards identifying the genetic determinants of metastasis in NSCLC. In this work, we have tried to identify genetic determinants of metastasis in NSCLC based on recurrence of mutations identified in NSCLC primary tumor and matched CNS metastases.

Currently, there are two models to explain appearance and systemic progression of primary tumors towards metastasis, viz, the linear progression model and the parallel progression model [9]. According to linear progression model, cells of primary tumor go through multiple rounds of mutation and selection in order to achieve capability of seeding metastasis. As the primary tumor clones expand, primary tumor cells disseminate from the site of primary tumor to seed metastasis at distant sites.

In other words, the cells of primary tumor accumulate genetic mutations and eventually acquire full metastatic potential after which they disseminate from the site of primary tumor and travel to an ectopic site. These disseminated tumor cells then seed metastasis. For example, mutations in TP53 gene are found at a significantly higher frequency in late stage primary tumors than in early stage (T1). Also, increase in tumor size is well known to correlate with higher frequency of metastasis (TNM staging system), further supporting the linear progression model. An extension of linear progression model is tumor self seeding [6], a phenomenon where circulating tumor cells can re-colonize the site of their origin and can seed more tumor clones thereby accelerating tumor growth and progression towards metastasis. Also, tumor cells disseminating from the site of metastasis are expected to have the capability of survival and clonal expansion at distant site, and can be a prominent source of further metastases, leading to a metastatic cascade [11].

Whereas, in the parallel progression model, tumor cells disseminate early from the site of primary tumor, and these early disseminated tumor cells colonize multiple distant sites and accumulate genetic changes independently of primary tumor, eventually acquiring capability to seed metastasis.

Some early studies [7,8] aimed at quantifying growth rare in human cancers concluded that metastatic tumors are too large to be seeded by primary tumor cells that have disseminated from late stage primary tumors. So, under the parallel progression model, greater genetic disparity is expected between primary tumors and matched metastases, than under the linear model. Unfortunately, neither of the two models is supported by direct or incontrovertible evidence and multiple indirect evidences exist in support for both models.

In another study that compared primary tumors with matched metastases [10], growth rate of metastatic tumors have been found to correlate with the growth rates of the matched primary tumors, and not with the site of metastasis, indicating that metastases may inherit many of their features from the source tissue. Thus, in the light of this evidence and to look for evidence in favor of either growth model, we chose to analyze pairs of matched primary tumors and ensuing metastases, than analyzing metastases matched by their distant sites.

Significant progress, however, has been made towards discerning the source of metastatic relapse clones in hematological tumors. Studies aimed at identifying clonal relationships between primary tumor and relapse clones in AML [1,2] revealed mutational selection of primary tumor and consequent evolution into relapse clones. These studies showed that the cells of founding clone, or a subclone of the founding clone survived initial therapy, accumulated additional mutations and went through clonal expansion to form a relapse. However, the same questions remain unanswered still for solid tumors.

In this work, we have used whole exome DNA sequence data derived from biopsies of primary tumor (Non Small-Cell Lung Cancer) and matched metastases detected in the central nervous system (CNS) (9 pairs of matched primary NSCLC tumor and CNS metastases) to analyse clonal relationships between primary tumor and matched metastasis. We also measured genetic divergence between primary tumors and matched metastasis and weighed the results in favor of either linear or parallel progression model. We analyzed mutation profiles of primary tumors and matched metastases to determine if metastatic clone was clonally derived from the dominant clone of primary tumor (thus supporting linear progression model), or if metastases are clonally

unrelated or genetically divergent with respect to primary tumor (supporting parallel progression model).

Any evidence in support of either of the two models would demand reconsideration of research strategies in prognosis and management of late stage cancer. Particularly, if the parallel progression model is found to be correct, then it would force researchers to reconsider current research approaches and clinical decision making paradigms that use primary tumors for providing markers for tailoring therapy.

Results and Discussion

I. Genetic determinants of metastasis

We used whole exome sequencing data of primary tumors and matched CNS metastases (9 patients). We detected single nucleotide polymorphisms using standard sequence alignment (Novoalign) and variant calling methods (VarScan/Samtools), details of which are in Methods section. In the absence of matched normal tissue, we used stringent filters for variant detection and applied additional filters to remove likely non-somatic variants. To filter out the non-somatic variants, we used common polymorphism databases like dbSNP, NHLBI's Exome Sequencing Project and 1000 Genomes project. Additionally, we also filtered out variants which are specific to the exome capture technology (viz, Agilent SureSelectV4, Agilent SureSelectV3 and Illumina TruSeq) or sequencing platform (Illumina HiSeq).

We identified the genetic mutations which were found to be recurrent in the primary tumor exomes. Recurrence in multiple primary tumor samples indicates a higher likelihood of being a driver mutation for either cancer or metastatic phenotype. We hypothesize that these mutations may be involved in a) promoting progression to, or maintenance of the cancer phenotype; or b) promoting the persistence of primary tumor cells as they escape in the vasculature, or their progression towards metastatic phenotype. Otherwise, these mutations may be just getting propagated as passenger mutations, which however would be less likely given that they are recurrent in multiple samples.

Validation of recurrent genetic mutations in a set of metastatic and non-metastatic biopsies

We obtained a set of 51 NSCLC primary tumor biopsies to perform validation of the recurrent genetic mutations identified in the set of whole exome data of primary tumor biopsies. 17 out of 51 of these validation set biopsies are from those patients which showed no evidence of clinical metastasis at the time of initial diagnosis and surgical resection of the primary tumor, but developed clinically evident relapse/metastatic lesions within 5 years of initial diagnosis, 15 were obtained from patients which showed metastasis in other patients, which the rest 19 were from patients which showed no clinical evidence of metastasis for 5 years or more after the initial diagnosis and surgical resection of the primary tumor.

We identified a set of 48 single nucleotide variants using a detailed selection process, which is highlighted in **figure 1**.

We performed validation for 48 single nucleotide variants selected previously in this set of primary tumor biopsies. Briefly, we designed primers approx. 70 bp upstream and downstream of the base of interest and PCR amplified the region. Amplicons were finally sequenced by Illumina MiSeq platform at 150X2 configuration. Reads were aligned back to hg19 whole genome reference, and variants were detected using VarScan using the same parameters are for initial variant detection for exome sequence data. The results of the validation pipeline are in **figure 2**.

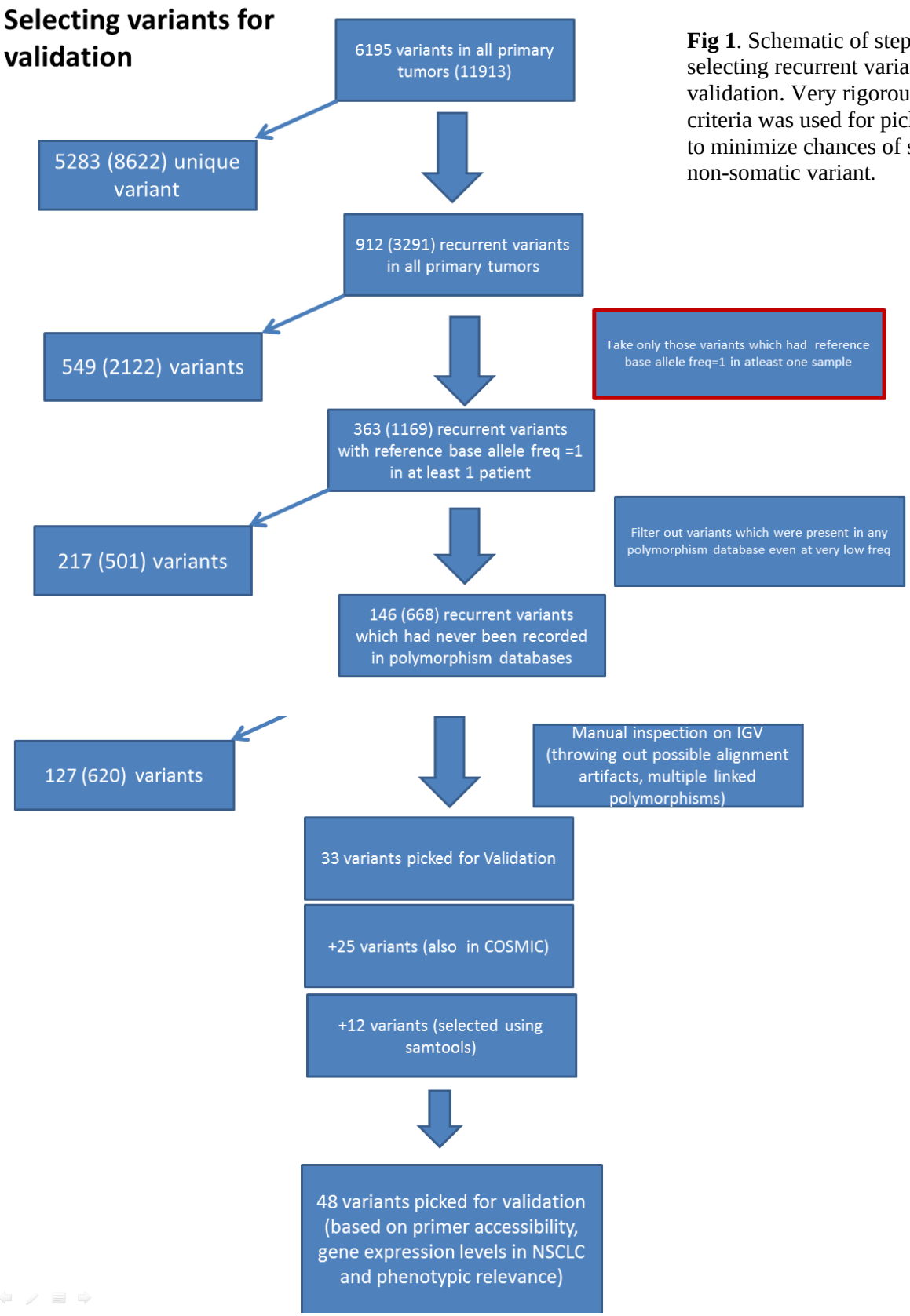


Fig 1. Schematic of steps involved in selecting recurrent variants for validation. Very rigorous selection criteria was used for picking variants to minimize chances of selecting a non-somatic variant.

				n=19	n=17	n=15	n=51
	Positive cases brain met	Positive cases no met	Positive cases other	No Met	Brain Met	Other Met	Average All Cases
IFT140	16	15	18	1.00	0.94	0.62	0.96
MLL3	15	10	13	0.67	0.88	0.45	0.75
COL4A6	12	13	17	0.87	0.71	0.59	0.82
FKBP9_H5	5	4	8	0.27	0.29	0.28	0.33
FOXD4L1_	5	2	4	0.13	0.29	0.14	0.22
FES	3	2	3	0.13	0.18	0.10	**0.16**
KRAS	3	1	2	0.07	0.18	0.07	0.12
CCDC37	3	1	0	0.07	0.18	0.00	0.08
PDLIM2	3	1	0	0.07	0.18	0.00	0.08
CR1L	2	3	1	0.20	0.12	0.03	0.12
TNFRSF10(	2	2	1	0.13	0.12	0.03	0.10
BAGE2	2	1	3	0.07	0.12	0.10	0.12
TP53	2	1	0	0.07	0.12	0.00	0.06
CAMKV	1	2	2	0.13	0.06	0.07	0.10
PCSK7	1	1	0	0.07	0.06	0.00	0.04
PRSS1	1	0	1	0.00	0.06	0.03	0.04
TAS2R46	1	0	1	0.00	0.06	0.03	0.04
DTX2_A2V	1	0	0	0.00	0.06	0.00	0.02
DEPTOR	1	0	0	0.00	0.06	0.00	0.02
SEC22B	0	3	5	0.20	0.00	0.17	0.16
TDP1	0	1	1	0.07	0.00	0.03	0.04
A2M	0	1	0	0.07	0.00	0.00	0.02
??	0	1	0	0.07	0.00	0.00	0.02
LGI4	0	1	0	0.07	0.00	0.00	0.02
TARBP1	0	0	1	0.00	0.00	0.03	0.02
HYAL3	0	0	1	0.00	0.00	0.03	0.02
FOXP4	0	0	1	0.00	0.00	0.03	0.02
PRDM10	0	0	1	0.00	0.00	0.03	0.02
HNRNPUL1	0	0	1	0.00	0.00	0.03	0.02
TRIL	0	0	0	0.00	0.00	0.00	0.00
CNTNAP4	0	0	0	0.00	0.00	0.00	0.00
MUC16	0	0	0	0.00	0.00	0.00	0.00

Figure 2. Recurrent mutations were validated in a set of 51 primary tumor biopsies. The genes bearing those mutations are listed in the descending order of recurrence in validation set. Several genes, such MLL3, FKbP9, FES, PDLIM2, are known to have functions related to metastasis, appeared to have recurrent mutations in validation set.

II. Clonal Relationships among Primary And Metastatic Biopsies

Clonal Divergence between Primary and Metastatic Biopsies

The variant allele frequencies of mutations found in primary were plotted against allele frequencies of variants found in metastatic tissues. We were able to broadly classify the patients

into two groups. The first group comprised of PT1, PT2, PT9 and PT13, which showed similar VAF plots, in the sense that there is a clear cluster of variants which are enriched in metastasis. The other group comprised of PT3, PT 6, PT10, and PT11, which showed no enrichment of any subclone in metastasis, and the primary tumor and metastatic variant allele frequencies correlate well. The scatter plots are shown in **figure 3**. We also performed K-means clustering on each of the 2-D scatter plots using k=6. We obtained distinct clusters comprising of variants enriched in metastasis in PT_13 data.

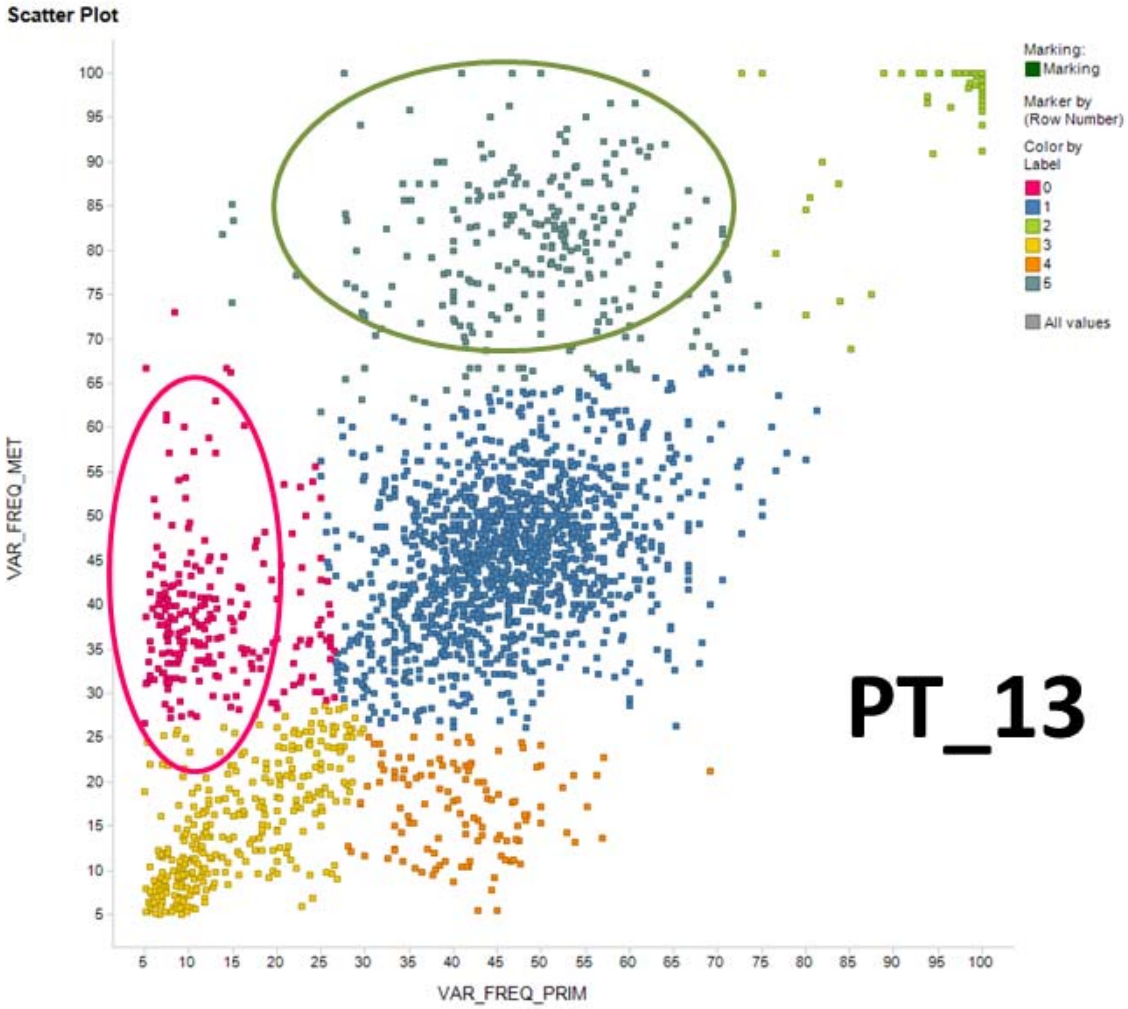


Figure 3-a. Variant allele frequencies of primary tumor were plotted against the allele frequencies seen in metastasis. K-means clustering with k=6 was applied, which resulted in distinct clusters (pink and green) representing variants enriched in metastasis.

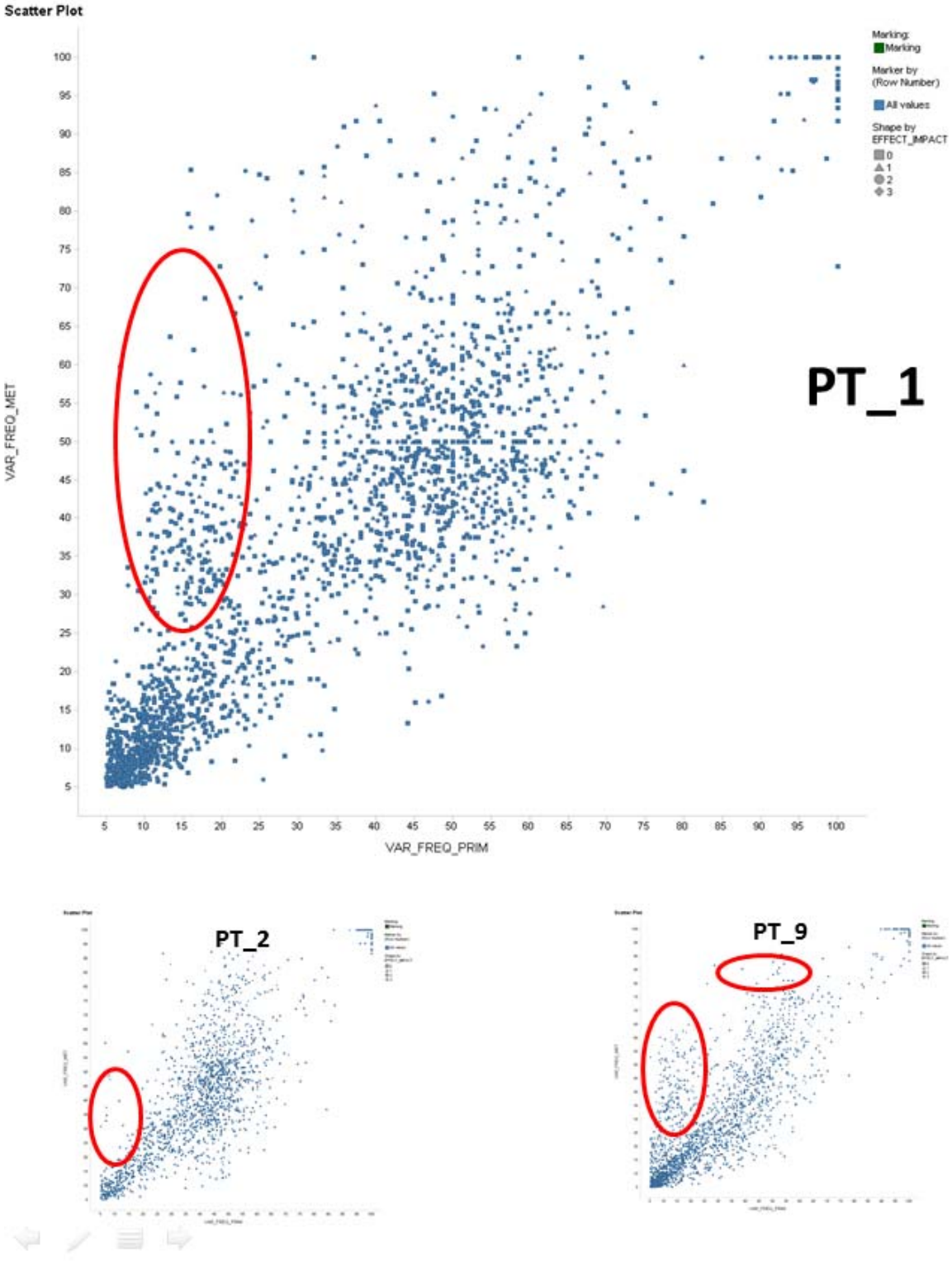


Figure 3-b. Variant allele frequencies of primary tumor were plotted against the allele frequencies seen in metastasis. Red ovals are used to highlight mutations enriched in metastasis.

We also calculated correlation (Pearson's) between primary and metastatic variant allele frequencies and straight line fit to the allele frequencies.

Correlation Coefficient (Pearson's R): A high correlation coefficient (CC) indicates that the primary tumor and metastasis are more closely related to each other, clonally. Conversely, a low correlation coefficient would indicate high clonal divergence between primary and metastasis tumors, which might be the result of more time spent between appearance of primary and metastatic tumors. However, a low correlation coefficient might also result from noise (false positive variants). Among pairs of primary tumors and metastases, the pairs having higher CC would mean that they are clonally more similar (metastatic tumor hasn't evolved much from the primary tumor) than the pairs having lower CC.

From the data, it appears that metastatic tumor is clonally quite similar to the primary tumor. A plausible explanation could be that primary tumor cells escaped into the vasculature and got deposited (and propagated) at the site of metastasis, particularly due to the highly similar the clonal composition of primary and metastatic clones in some patients. From the distribution of allele frequencies of metastasis, it also seems like metastatic clone has not resulted from clonal expansion of a primary tumor cell.

We also performed a straight line fit to the distribution of allele frequencies in primary v/s metastatic biopsies. A high slope (>0.45) would indicate that variants are enriched in metastasis compared to primary, overall. Conversely, a low slope (<0.45) would indicate that variants are under-represented (or lost from metastasis) in metastasis compared to primary. The results of correlation analysis and straight line fit are tabulated in **figure 4**.

	Correlation (Pearson's R)	Number of Rare Variants	Slope (Straight line fit)	Intercept
PT_1	0.5818387	4204	0.626	11.49
PT_2	0.7446795	3785	0.7228	14.3976
PT_3	0.7275213	4735	0.6919	10.399
PT_6	0.7433719	3556	0.7908	11.0838
PT_9	0.5748225	4145	0.6112	18.4449
PT_10	0.6712747	4611	0.6447	9.3206
PT_11	0.7363403	5199	0.7207	4.7112
PT_12	0.6340622	5069	0.6271	5.6524
PT_13	0.480376	3889	0.4949	20.9449

Figure 4: Correlation coefficients (Pearson's R) between allele frequencies of Expected Somatic variants in primary and metastatic biopsies along with the slope of the straight line fit are tabulated.

PT_1: The scatter plots shows that variants are loosely scattered.

PT_2: There appears to be a cluster of variants which are enriched in metastasis, and another cluster which is under-represented in metastasis. PT_2 also has the highest correlation coefficient and the second highest slope (of straight line fit). From the high slope, we can saw that variants are enriched in metastasis compared to primary, inspite of a tight correlation between primary and metastasis.

PT_3, PT_6 and PT_11: Similar to PT2, we see a good correlation between the variant allele frequencies of primary and metastasis, and a high slope, indicating that allele frequencies are higher for metastasis, compared to primary (thus proving that, in general, variants have got enriched in metastasis compared to primary).

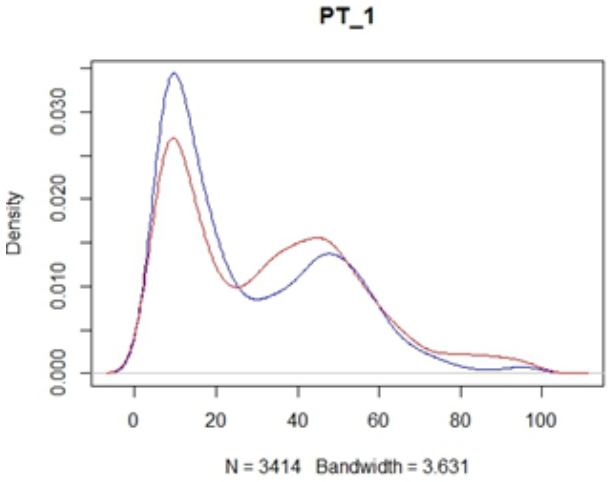
PT_13: Here, we see a low correlation between allele frequencies of primary and metastasis. It is also evident from the scatter plots that there are several clusters of mutations, one particular cluster comprising of mutations that are enriched in metastasis, while another cluster comprising of mutations which are under-represented in metastasis. So, the metastasis tumor is clonally distant from the primary tumor.

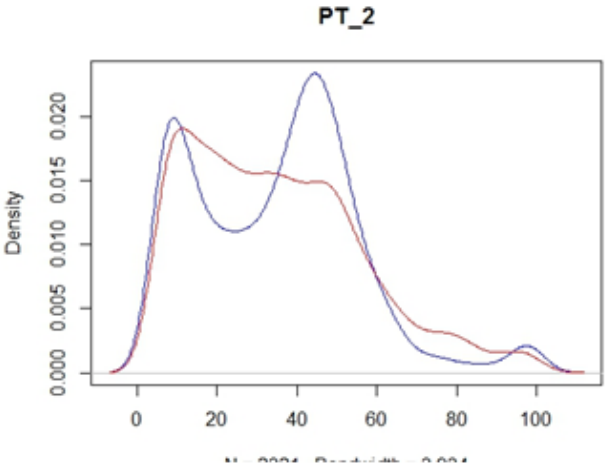
Clonal architecture of primary and metastatic tumors

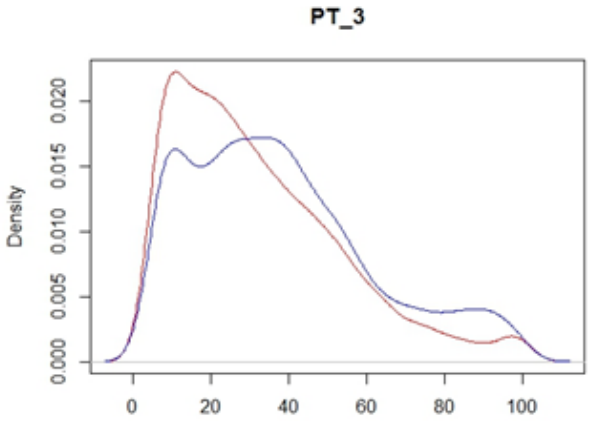
To disseminate the clonal architecture of primary and metastatic tumors, we performed kernel density analysis on the variant allele frequencies of primary and metastatic tumor biopsies. The kernel density plots (shown in **figure 5**) were generated using 'density' function in R; more details are in the methods section.

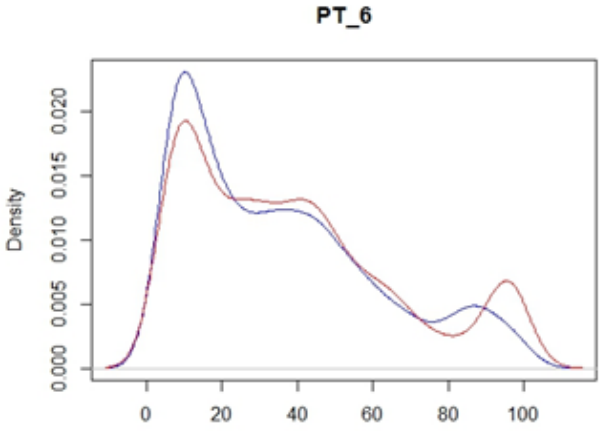
The density plots reveal two major density peaks per biopsy, one around the allele frequency of 10% and the other one around the allele frequency of 45%. Considering the amount of noise and non-somatic mutations present in the dataset, we believe that the peak near AF of 45% is mostly composed of non-somatic (germline mutations), though it might also have somatic mutations. The other peak, which is around 10% AF is expected to be mostly composed of somatic mutations. Considering mutations are in heterozygous state, this also gives us an estimate of the tumor purity (~20%).

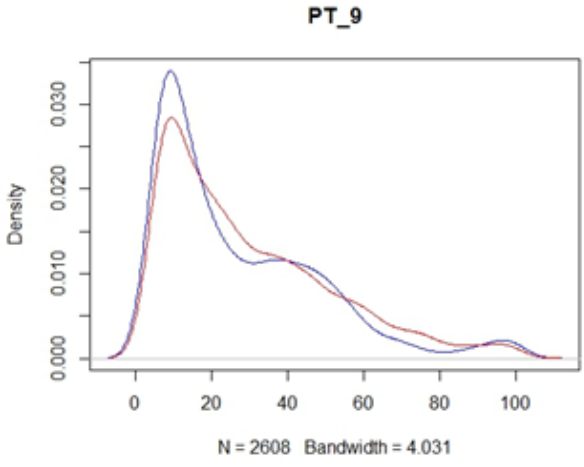
An alternative explanation could be that the peak at 10% allele frequency is the secondary clone, while the peak at 45% allele frequency is the primary (or dominant) clone. As per this estimate, the tumor purity would be around 90%.

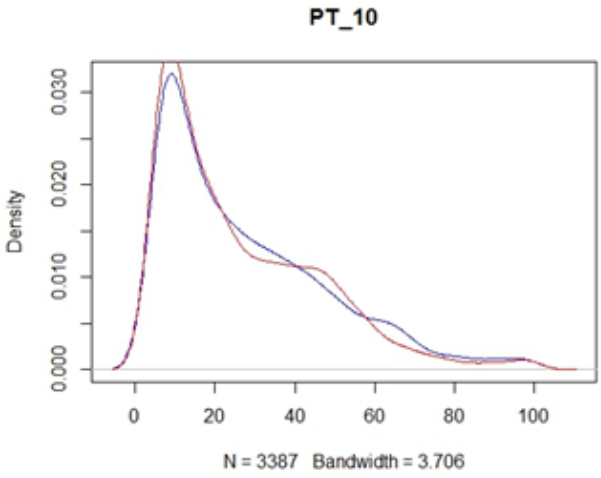

PT_1
Density
0.000
0.010
0.020
0.030
0
20
40
60
80
100
N = 3414 Bandwidth = 3.631


PT_2
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
100


PT_3
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
100


PT_6
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
100


PT_9
Density
0.000
0.010
0.020
0.030
0
20
40
60
80
100
N = 2608 Bandwidth = 4.031


PT_10
Density
0.000
0.010
0.020
0.030
0
20
40
60
80
100
N = 3387 Bandwidth = 3.706

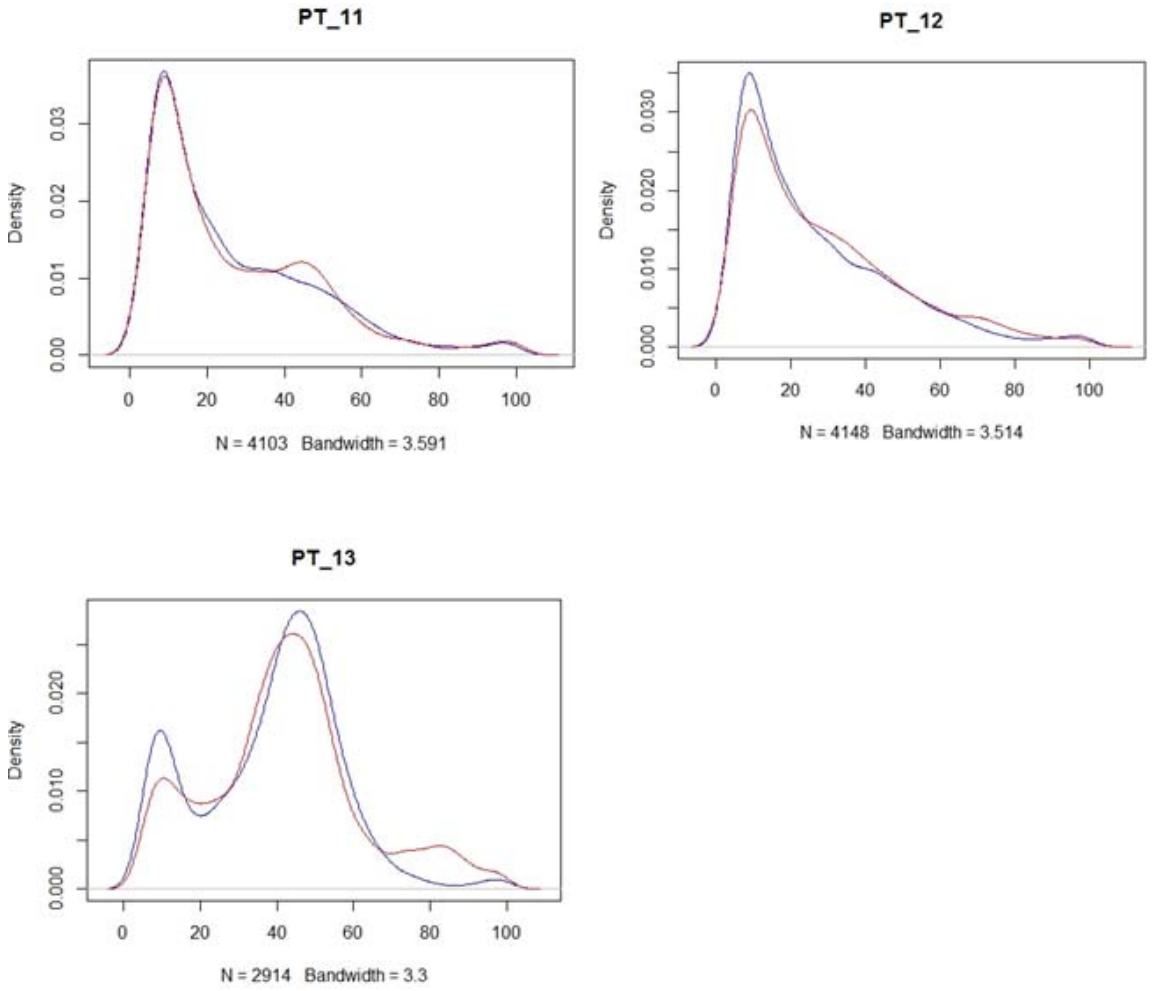


Figure 5: Kernel density plots of variant allele frequencies in respective patients biopsies were generated using density function in R. Blue curve indicates primary tumor variants, while the red curve represents metastatic variants.

III. Sensitivity / Specificity Analysis

We tried to determine if we could get an estimate of the amount of noise (non-somatic mutations) in the dataset. For this analysis, we used whole exome data from primary tumor and matched normal tissues (solid tumor, gastric cancer) processed on Agilent SureSelect V4 platform. We derived the set of True Somatic deleterious variants by subtracting the variants

found in normal from the variants found in tumor tissue. We also derived a set of Expected Somatic deleterious variants by taking the deleterious variant set of tumor sample and subtracting the set of mutations which also happened to be in the Common polymorphism databases (consisting of variants from dbSNP, EVS, Internal Controls; more details are in Methods section).

Briefly,

Expected Somatic deleterious = Tumor sample deleterious mutations – Common variants

True Somatic deleterious = Tumor sample deleterious mutations – (Matched) Normal sample mutations

We overlapped the sets of variants obtained as above to determine true positive variants, false positive and false negative variants.

True Positive: Deleterious variants which are present in both Expected Somatic Deleterious set and True Somatic Deleterious Set

False Negatives: False Negatives are variants which were true somatic (present in cancer, absent in normal), but absent from Expected Somatic list. To determine why these variants weren't present in "Expected Somatic" group, we overlapped the False negative variants with the common polymorphisms set and found that all False Negative variants overlap 100% with common polymorphisms set. However, we did not determine whether these False Negative variants were undetected in normal tissue due to lack of coverage, or they were actually absent from normal tissue (hence truly somatic). If they were actually absent from normal, and hence are true somatic variants, then it would be interesting to determine why these variants appear in large frequencies in polymorphism databases.

False Positives: These variants are expected to be germline polymorphisms, meaning that they are present in both tumor as well as normal. So, they were excluded from True Somatic, but still were present in likely somatic since these variants are still not present in any polymorphism databases.

The venn diagrams below (**figure 6**) describe the overlap of mutations between true somatic and expected somatic.

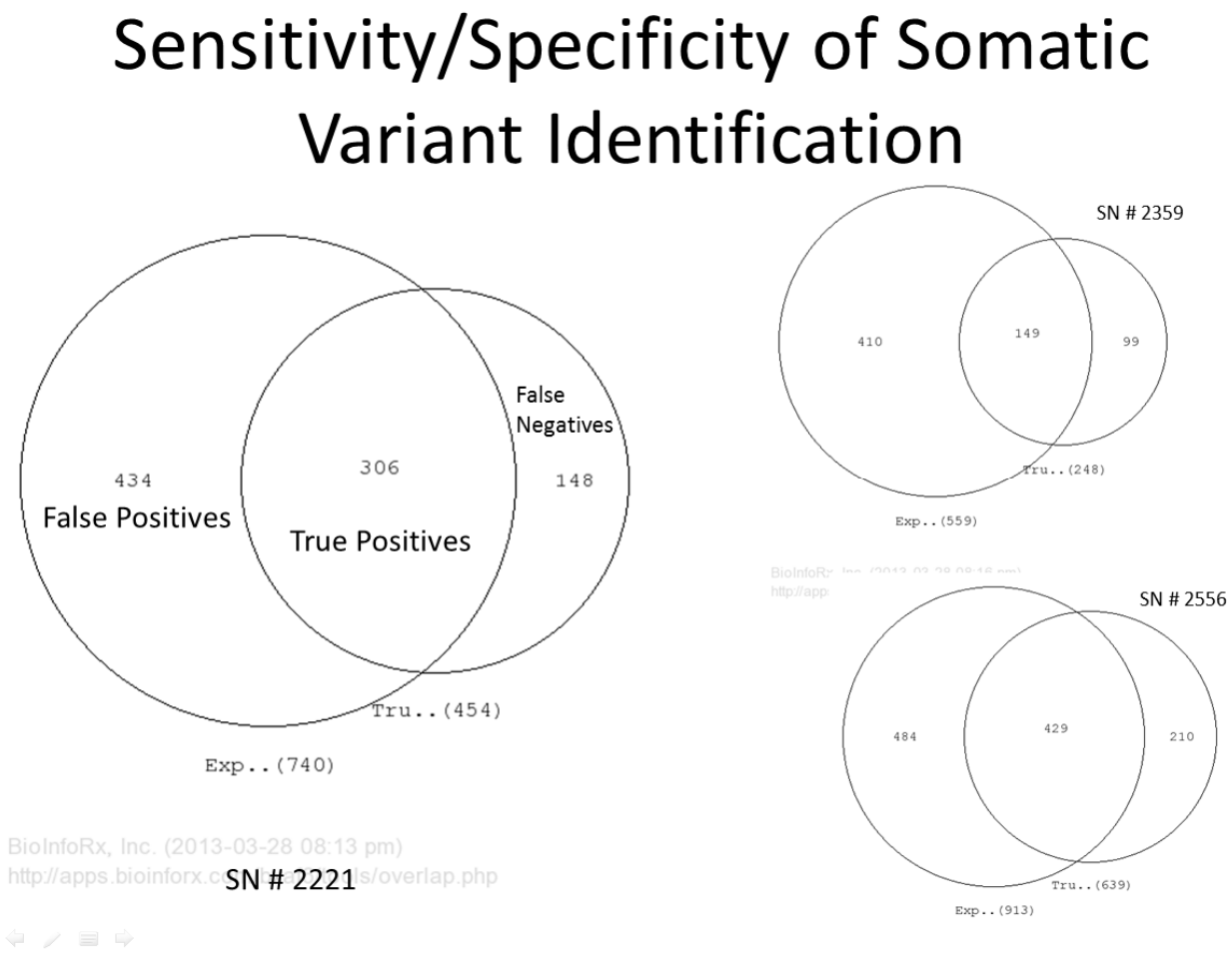


Figure 6: Set overlap of Expected somatic deleterious variants with True Somatic variants is depicted using Venn Diagrams.

Moreover, we calculated the recurrence among true positives and false positive variants. While picking the variants for validation (from 9 pairs of primary-metastatic pairs), we assumed that a false positive variant is unlikely to be recurrent in multiple patients, unless it is a result of

systematic or platform specific error. So, we searched for recurrent variants among the set of True Positive variants (variants which are present in True Somatic, and also present in Expected Somatic), and False Positive variants (variants which were present in Expected Somatic, but absent in True Somatic). We found a set of 31 recurrent true positive variants (present in True positive set of 2 or more samples (sample size = 4 tumor normal pairs)) and 53 False Positive variants (present in False Positive Set of 2 or more samples in a sample size of 4 tumor normal pairs). Evidently, we got almost twice the number of False Positives recurrent variants as True Positives recurrent variants, further highlighting that the variants we picked for validation might have a high content of noise. Also, among the variants we picked for validation, only 1/3 are expected to be True somatic recurrent variants.

Pitfalls:

1. Variants which are recurrent in multiple patients might have a wide range of allele frequencies ($0.05<AF<0.5$). Eg, assuming all acquired mutations are heterozygous, an allele frequency of 0.5 would indicate that the variant is present in almost all the cells of the biopsy, whereas an allele frequency of 0.05 would indicate that the variant is present in utmost 10% of the cells of the biopsy. Primary tumors display a significant degree of mutational heterogeneity among themselves [21], and single tumor-biopsy samples have been shown to under-represent the mutational and genomic heterogeneity within primary tumors [22]. So, even though a variant shows high variability in its allele frequency between samples, it may be representing the dominant clone of the primary tumor.

2. Also, since single biopsies of solid tumors have been known to under-represent the actual amount of genomic heterogeneity [22], we expect to have a lower recall rate (low sensitivity) for recurrent mutations.

3. Also, these biopsies have been derived from patients who have undergone different non-specific anti-cancer therapeutic regimes for different time-periods. Though the effect of different anti-cancer therapies on overall mutation profile or overall genomic landscape is unknown, we expect that the effect of such differences to be non-systemic and small and insignificant.

Methods

I. Sequence Alignment and Variant Detection

1. Alignments were done using hg19 reference genome and Novoalign (version -2.08.02).

Novoalign Parameters Used:
-o SAM -r none -l 30 -e 100 -i 230 140 -a AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA -H -c 12
Parameters guide:
-o Format
-r Strategy for reporting repeats (None, Random, All, Exhaustive)
-l minimum length for good quality sequence
-e Number of alignments for a single read (def = 999)
-i Approximate fragment length (for paired end reads)
-a trimming of adapter sequences
-H hardclipping of the low quality seqments
-c number of cores to use

2. Next, we used samtools to sort and merge all the bam files for a particular sample. Samtools version 0.1.14 was used.
3. Then, we used Picard-MarkDuplicates.jar to remove duplicate reads (Picard version 1.67).
4. Then, we used intersectBed to create a bam file having reads which overlap with the exon capture coordinates (IntersectBed version 2.16.2).
5. Then, we used samtools –flagstat method to calculate read alignment statistics.

Variant Detection

Single nucleotide variant detection was performed using VarScan (mpileup2snp) using the following params.
(--min-reads2 5 --min-var-freq 0.05 --p-value 0.05)
Filters:
#Minimum base quality at position to count a read =15
#P-value threshold = 0.05
#Minimum supporting reads at a position to call variants = 5
#Minimum variant allele freq = 0.05

Variants were annotated for synonymous and non-synonymous status and mutation effect using snpEff and further annotations regarding status in 1000G, dbSNP and other databases using dbNSFP (27) and SnpSIFT (28, 29).

II. Validation pipeline

1. Alignments were done using hg19 reference genome and Novoalign (V2.08.03).

Novoalign Parameters Used:
-o SAM -r None -l 30 -e 100 -i PE 95-300 -a TGTAGAACCATGTCGTCAGTGTGTGCTCATGTATCTCGTATGCCGTCTTCG AGACCAAGTCTCTGCTACCGTGTAGATCTCGGTGGTCGCCGTATCATT -H -c 8
Parameters guide:
-o Format
-r Strategy for reporting repeats (None, Random, All, Exhaustive)
-l minimum length for good quality sequence
-e Number of alignments for a single read (def = 999)
-i Approximate fragment length (for paired end reads)
-a trimming of adapter sequences
-H hardclipping of the low quality seqments
-c number of cores to use

I had also created a custom reference, which contained the amplicon sequences (+100 bases upstream and downstream), to align the reads. While using the custom reference, 0.5-2% more reads aligned to the reference. The alignment success rate with hg19 reference index was 89-94%. However, since the validation assay is based on whole genome (the primers can anneal anywhere their complimentary sequence exists), I chose to use the whole genome reference (hg19) rather than the custom reference.

2. Next, I used samtools to sort and index all the bam. Samtools version 0.1.18 was used.
3. Next, I used samtools' mpileup feature to create pileup files for each bam alignment.
4. Next, I used VarScan's mpileup2snp to get variant calls from respective mpileup files**. VarScan parameters** used :
   VarScan mpileup2snp --min-reads2 5 --min-var-freq 0.05 --p-value 0.05
   Other parameters were applied by default [in square parenthesis]

--min-coverage =8	Minimum read depth at a position to make a call [8]
--min-avg-qual =15	Minimum base quality at a position to count a read [15]
--strand-filter =1	Ignore variants with >90% support on one strand [1]
--output-vcf =1	If set to 1, outputs in VCF format
--variants =0	Report only variant (SNP/indel) positions (mpileup2cns only)[0]

III. Variant Allele Frequency Plots (2D scatter plots)

1. Sequence alignment was done using Novoalign and VarScan was used for variant detection.
2. snpEFF and SnpSift were used to annotate the output of VarScan.
3. I developed a set of ***common polymorphisms*** by picking variants from polymorphism databases like dbSNP, EVS, and also from internal database of variants which were found to be common to the platform used for exon capture (Agilent V3/V4 / Illumina TruSeq etc.). The selection criteria for including variants in the set of ***common polymorphisms*** is as follows***:***
   {{dbSNP } OR {EVS} OR {Internal Control}}

dbSNP: Global Minor Alllele Frequency >= 0.01 OR G5A tag is present. dbSNP derives its Global Minor Allele Frequency stats from 1000Genomes Project phase 1 (genotype data from 629 individuals, released in the 11-23-2010 dataset). A variant is tagged with G5A tag when the minor allele frequency >5% in each and all populations.

OR

EVS: A variant is considered as a common variant when:

Total Samples Covered >=100 (meaning that the variant should be present in atleast 100 samples).

AND

Average sample read depth >= 10 (meaning that the variant location is covered by an average of 10 bases in all samples combined).

AND

Minor allele Frequency >= 3% (If a variant is present in more than 5 individuals out of 100 in heterozygous, then the allele frequency would be 0.03 (or 3%). So I used a minor allele frequency threshold of 3%).

OR

Internal Controls:

I took pooled the variant calls from 100 samples processed on Illumina TruSeq platform and Agilent V3 platform, respectively (kindly provided by GTAC/Paul Cliften). If a variant occurs 5 times or more in the pool of variant calls belonging to a particular platform, then I consider that variant as a platform specific common variant.

Finally, I pooled all the common polymorphisms from dbSNP, EVS and internal controls into one '***common polymorphism'*** dataset. I used this dataset to filter out any potentially common (non-somatic) mutation found in our samples.

So, after filtering out common polymorphisms, I made a list of (expected somatic) rare polymorphisms for each patient (pair of primary and metastasis). The allele frequencies of variants in primary tumor were plotted against the respective variants' allele frequencies in metastatic tumor in a scatter plot using SpotFire. Correlation coefficient and fitting of linear model (straight line fit) were done in R.

IV. Clonality Analysis

Variants were obtained as before, using Novoalign and VarScan. However, for this analysis, all Expected rare mutations were considered (synonymous and non-synonymous). Specifically, the Expected rare mutation set was derived in the following way.

Expected Rare = Tumor (Primary or Metastatic) variants – Common Variants

Kernel density Analysis was done using *density* function in R.

V. Sensitivity Analysis

We picked whole exome data from solid tissue tumor samples (gastric cancer). The samples were processed on Agilent V4 platform by Genome Technology Access Center, who also

processed the 9 pairs of primary and matched metastatic tumor biopsies. Sequence alignment and variant detection were done using exactly the same tools and parameters as for the primary-metastatic pairs.

References

1. Eheman C. et al. Annual Report to the Nation on the Status of Cancer, 1975-2008, Featuring Cancers Associated With Excess Weight and Lack of Sufficient Physical Activity. Cancer. 2012, 118(9) 2338-66

2.Fidler IJ. The organ microenvironment and cancer metastasis. Differentiation. 2002;70(9-10), 498-505.

3. Kreso A, et al. Variable Clonal Repopulation Dynamics Influence Chemotherapy Response in Colorectal Cancer. Science. 2012; 339(6119):543-8

4.Comprehensive genomic characterization of squamous cell lung cancers. Cancer Genome Atlas Research Network, Nature. 2012;489(7417):519-25

5. Jeong-Sun Seo, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research. 2012; 22(11):2109-19

6. Kim MY, et al. Tumor Self-Seeding by Circulating Cancer Cells. Cell. 2009;139(7):1315-26

7. Collins VP, et al, Observations on growth rates of human tumors. Am. J. Roentgenol. Radium Ther. Nucl. Med. 1956;76, 988–1000.

8. Friberg S et al. On the growth rates of human malignant tumors: implications for medical decision making. J. Surg. Oncol. 1997;65, 284–297.

9. Klein CA. Parallel progression of primary tumours and metastases. Nat Rev Cancer. 2009; 9(4): 302-12

10.Kusama S, et al. The gross rates of growth of human mammary carcinoma. Cancer. 1972;30,594–599.

11. Bross I D, et al. Do generalized metastases occur directly from the primary? J. Chronic Dis. 1975;28, 149–159

12. Husemann Y, et al. Systemic spread is an early step in breast cancer. Cancer Cell. 2008;13,58–68.

13. Ding L, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012; 481(7382):506-10

14. Minn AJ, et al. Genes that mediate breast cancer metastasis to lung. Nature.2005;436(7050):518-24.

15. Kang Y, et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell. 2003;3(6):537-49.

16. Bos PD, et al. Genes that mediate breast cancer metastasis to the brain. Nature. 2009; 59(7249):1005-9.

17. Harbour JW, et al. Frequent mutation of BAP1 in metastasizing uveal melanomas. Science. 2010;330(6009):1410-3.

18. Vultur A, et al. BRAF inhibitor unveils its potential against advanced melanoma. Cancer Cell. 2010;18(4):301-2.

19. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multi-region sequencing. N Engl J Med. 2012; 366(10):883-92.

20. Stoecklein NH, et al. Genetic disparity between primary tumours, disseminated tumour cells, and manifest metastasis. Int J Cancer. 2010;126(3):589-98.

21Torres L, et al. Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast Cancer Res Treat. 2007; 102(2):143-55

22. Nguyen DX, et al. Genetic determinants of cancer metastasis. Nat Rev Genet. 2007; 8(5):341-52.

23. Walter MJ, et al. Clonal Architecture of Secondary Acute Myeloid Leukemia. NEJM. 2012; 366(12):1090-8

24. Slamon DJ, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344(11):783-92.

26. Liu X, et al. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation. 2011;32(8), 894–9.

26. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.

27. Cingolani P, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics. 2012;3:35