PLOS | GENETICS

RESEARCH ARTICLE

Preferential Allele Expression Analysis Identifies Shared Germline and Somatic Driver Genes in Advanced Ovarian Cancer

Najeeb M. Halabi[1], Alejandra Martinez[2], Halema Al-Farsi[1], Eliane Mery[3], Laurence Puydenus[3], Pascal Pujol[4], Hanif G. Khalak[5], Cameron McLurcan[6], Gwenael Ferron[2], Denis Querleu[2], Iman Al-Azwani[7], Eman Al-Dous[7], Yasmin A. Mohamoud[7], Joel A. Malek[1,7], Arash Rafii[1,8]*

1 Department of Genetic Medicine, Weill-Cornell Medical College, New York, United States of America, **2** Surgery Department, Institute Claudius Regaud, Toulouse, France, **3** Pathology Department, Institute Claudius Regaud, Toulouse, France, **4** Oncogenetics, Centre Hospitalier Regional Universitaire de Montpellier, Montpellier, France, **5** Advanced Computing, Weill-Cornell Medical College in Qatar, Doha, Qatar, **6** Biosciences Department, University of Birmingham, Birmingham, United Kingdom, **7** Genomics Core, Weill-Cornell Medical in Qatar, Doha, Qatar, **8** Stem Cells and Microenvironment Laboratory, Weill-Cornell Medical College in Qatar, Doha, Qatar

* jat2021@qatar-med.cornell.edu

Citation: Halabi NM, Martinez A, Al-Farsi H, Mery E, Puydenus L, Pujol P, et al. (2016) Preferential Allele Expression Analysis Identifies Shared Germline and Somatic Driver Genes in Advanced Ovarian Cancer. PLoS Genet 12(1): e1005755. doi:10.1371/journal.pgen.1005755

Editor: Elizabeth Swisher, University of Washington, UNITED STATES

Received: June 6, 2014

Accepted: November 30, 2015

Published: January 6, 2016

Data Availability Statement: Primary data consisting of non-mitochondrial RNA and exome sequencing data are available from the GEO database (accession number GSE75935).

Funding: This work was supported by a grant from the Qatar National Research Fund (4-640-1-096). 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.

Abstract

Identifying genes where a variant allele is preferentially expressed in tumors could lead to a better understanding of cancer biology and optimization of targeted therapy. However, tumor sample heterogeneity complicates standard approaches for detecting preferential allele expression. We therefore developed a novel approach combining genome and transcriptome sequencing data from the same sample that corrects for sample heterogeneity and identifies significant preferentially expressed alleles. We applied this analysis to epithelial ovarian cancer samples consisting of matched primary ovary and peritoneum and lymph node metastasis. We find that preferentially expressed variant alleles include germline and somatic variants, are shared at a relatively high frequency between patients, and are in gene networks known to be involved in cancer processes. Analysis at a patient level identifies patient-specific preferentially expressed alleles in genes that are targets for known drugs. Analysis at a site level identifies patterns of site specific preferential allele expression with similar pathways being impacted in the primary and metastasis sites. We conclude that genes with preferentially expressed variant alleles can act as cancer drivers and that targeting those genes could lead to new therapeutic strategies.

Author Summary

Identifying genes that contribute to cancer biology is complicated partly because cancers can have dozens of somatic mutations and thousands of germline variants. Somatic mutations are gene variants that arise after conception in an organism while germline variants are gene variants present at conception in an organism. Most methods to identify cancer

drivers have focused on determining somatic mutations. In this study we attempt to identify, from a tumor sample, important germline and somatic variants by determining if a variant is expressed (made into RNA) more than expected from the amount of the variant in the genome. The preferred expression of a variant could benefit cancer cells. When applying our analysis to ovarian cancer samples we found that despite the apparent heterogeneity, different patients frequently share the same genes with preferentially expressed variants. These genes in many cases are known to affect cancer processes such as DNA repair, cell adhesion and cell signaling and are targetable with known drugs. We therefore conclude that our analysis can identify germline and somatic gene variants that contribute to cancer biology and can potentially guide individualized therapies.

Introduction

Identifying genes contributing to tumor biology (driver genes) underlies the design of targeted therapies. The advent of large-scale tumor sequencing in 2006 [1] followed by integrated multi-dimensional TCGA studies [2] brought a wealth of molecular data in different cancers at the somatic mutation, gene expression and copy number variation levels. One surprising result has been the observation that in most studied cancers, there are large differences in somatic mutations in patients. For example, in the case of the TCGA ovarian cancer study [3], there were ~10,000 somatic mutations among 316 patients with only *TP53* found mutated in the majority (96%) of patients. Every other gene was found to be mutated at low frequencies. This heterogeneity was also seen in more recent multi-site ovarian cancer studies [4, 5]. Contrary to the heterogeneity observed at the somatic mutation level, gene expression profiles are more homogeneous with distinct gene expression clusters observed both in the TCGA study and other studies [5, 6].

While both somatic mutation and gene expression studies have yielded large insights into tumor biology, they have several limitations in uncovering driver genes. Somatic mutations do not identify germline variants that contribute to tumor biology, require large patient cohorts, make assumptions about the background mutation rate and have turned out be very heterogeneous. Gene expression array studies, though they uncover sets of genes that correlate with prognosis, do not inform about significant or causative genes and do not indicate whether a mutated form of the gene is being expressed.

One approach to address some of these limitations is to identify preferentially expressed variants of gene. If a specific variant is expressed and if that expression is at a significantly greater or lower level than expected, then this could indicate selection for or against that variant and imply that the gene is playing an important role in the tumor. This approach, called previously allelic expression bias analysis [7–9], typically determines significance if the expression allele fraction deviates from 0.5 (the expected non-biased allele expression at heterozygous positions). Allelic expression bias analysis however is difficult to apply to patient tumor samples because samples are often a mixture of normal and tumor cells with tumor cells themselves being heterogeneous with large copy number changes. Therefore, assumptions that an allele has biased expression if it differs significantly from a specific allele fraction are not justified.

However, combined genome and transcriptome sequencing (CGTS) as described here makes it possible to directly assess the genomic content of the sample at all sites and therefore determine if allelic expression is significantly different than expected from the genome content. A resulting advantage of CGTS is also a large reduction of alignment bias [10].

We applied CGTS to primary and metastatic epithelial ovarian cancer samples. Ovarian cancer is the most aggressive gynecological malignancy in developed countries. Ovarian cancer had been thought to arise from ovarian epithelial cells but more recent studies [11–15] have shown that at least 50% of ovarian tumors most likely arise from the fallopian tube. Regardless of origin, the ovaries and abdominal cavity are mostly impacted with more than 70% of patients diagnosed with disease spread throughout the abdominal cavity. These patients have a 30% five year survival rate [16]. Ovarian cancer dissemination most commonly occurs through the intraperitoneal route, followed by lymphatic invasion [17–19]. While 80% of patients with advanced epithelial ovarian cancer initially respond to primary treatment, most recur with a drug resistant phenotype. A subset of patients with clinically and pathologically indistinguishable disease develops a less aggressive disease and may survive much longer. Consequently, patients have biologically different diseases [20]. Furthermore, disease in the same patient can be biologically different according to tumor location and according to tumor temporal variations [21, 22]. Intratumor heterogeneity within the same patient is clinically relevant because status of predictive biomarkers could be used to adapt treatment decisions [23].

Given that patients mostly present with metastasis and that lethality is high at this stage, we include in our study matched primary and metastasis samples from three ovarian cancer patients. The primary sample is from the ovary and the metastasis samples are from the peritoneum and lymph node, the two most frequent metastasis sites [24]. We identify a biologically interesting set of shared significant preferentially expressed alleles, predict patient-specific drug targets and identify site-specific genes that may contribute to specific biology. These results suggest that preferentially expressed variant analysis can identify potential cancer drivers.

Results

Identifying Significant Preferentially Expressed Alleles

Our approach to identify causative mutations using both RNA and exome sequencing data was motivated by an observation made about the variants present in *TP53* in three patients. In our dataset, each patient had a different high or moderate impact variant in *TP53*. High and moderate impact variants are those changing the protein sequence or that affect a splice site. When comparing the RNA and exome sequencing data of those variants, we observed that while the variant alleles were present in both the RNA and exome data, the TP53 variant allele was overrepresented in the RNA sequencing data when compared to the exome sequencing data (Fig 1). Given the known centrality of *TP53* variants in cancers (TP53 somatic mutations have been observed in 96% of ovarian cancer samples [3]), this suggested to us that there could be a selection process occurring in these patients where variant *TP53* is overexpressed or normal *TP53* is suppressed or both.

To understand if this preferential allele expression exists in other genes, we performed a systematic analysis at all sites where RNA or exome sequencing data exists as outlined in Fig 2. We performed both an analysis for somatic mutation preferential expression and a combined germline/somatic variant preferential expression. In both analysis, we identify variant positions and calculate the reference allele fraction which is the number of reference allele reads divided by the total number of reads. Using the RNA sequencing data of a sample that correspond to the exome variant position, we also determine the RNA reference allele fraction. Then, we calculate the significance of the difference as described more fully in the methods. Throughout this study, we adopt a convention where the reference allele fraction is the standard and the alternate allele is the variant or mutant allele. Differences between the RNA and exome allele

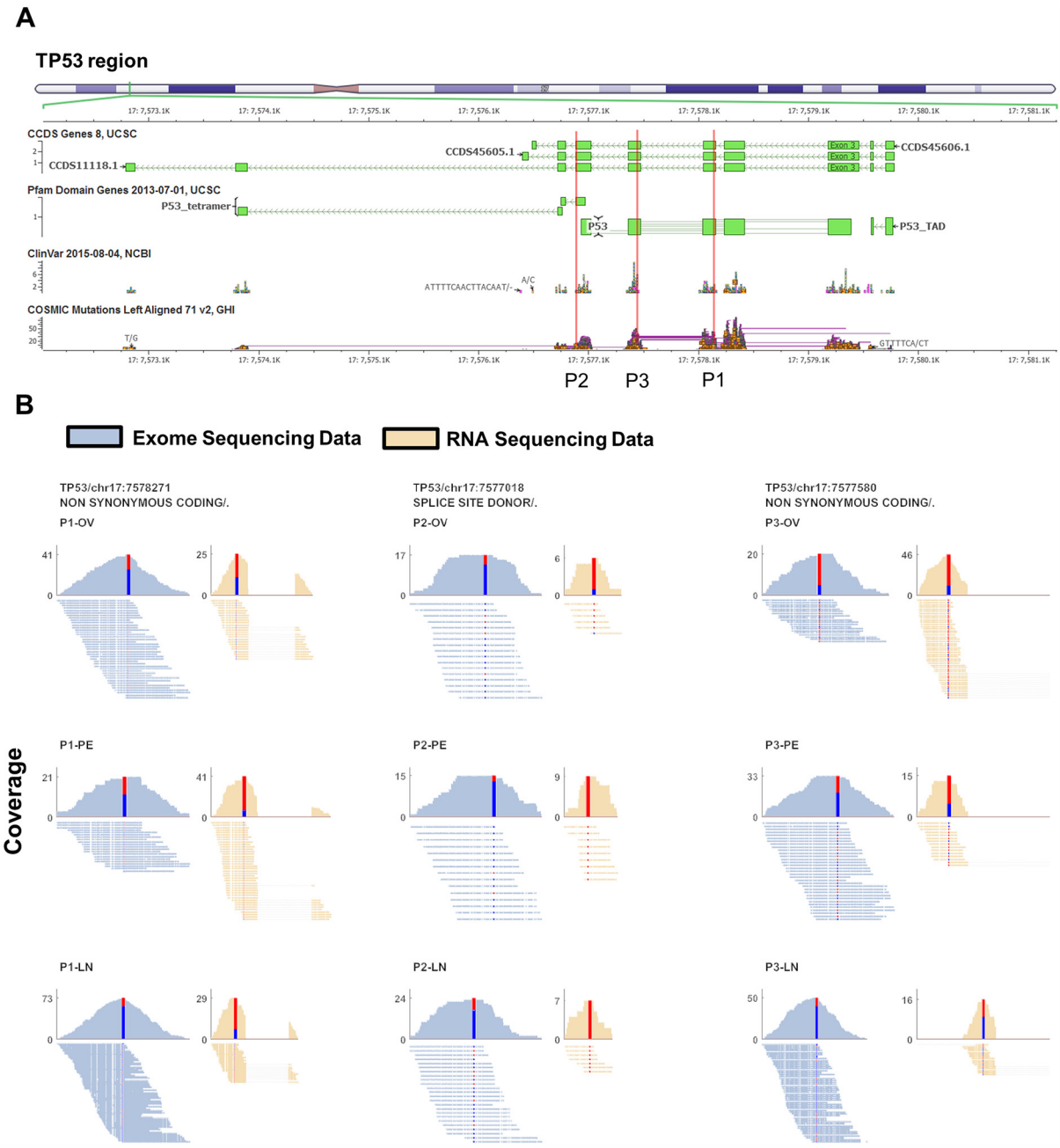


Fig 1. TP53 significant preferential allele expression. We observed in three different patients (P1, P2, P3) alleles in TP53 where one allele is significantly preferentially expressed. A) The allele positions are shown on a schematic view of the TP53 gene (red line) for all three patients. B) Alignment and coverage plots of the reads spanning the preferentially expressed alleles for both exome (blue) and RNA (brown) sequencing data. The patient and site is indicated on each plot. The histograms indicate coverage across the site with the coverage indicated by the y-axis number. The histogram central stacked bar plot shows the number of reference allele reads in blue and the alternate allele reads in red. The alignment plot shows the individual reads spanning the preferentially expressed alleles. For the RNA sequencing data, the dashed line indicate gaps corresponding to a spliced region.

doi:10.1371/journal.pgen.1005755.g001

fractions are calculated as RNA reference allele fraction—Exome reference allele fraction (abbreviated as RAD). Therefore, if the RNA reference allele fraction is 0.5 and the exome allele reference allele fraction is 0.9, this means that there is preferential expression of the variant allele with an allele fraction difference (RAD) of -0.4.

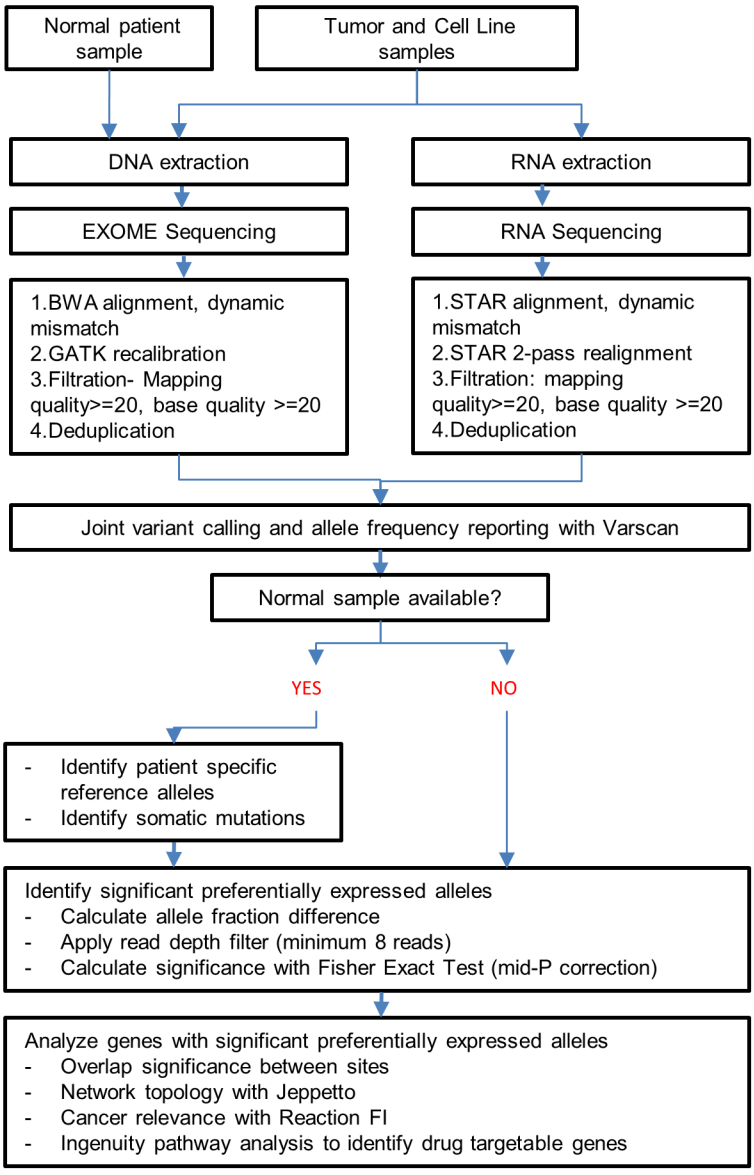


Fig 2. Schematic for global analysis of preferential allelic expression showing the steps followed for alignment, filtration, quality control and analysis steps.

doi:10.1371/journal.pgen.1005755.g002

The analysis of all somatic mutations (in two patients where non-cancer tissue is available) reveals mutated genes where the mutated allele is preferentially expressed (Fig 3). TP53 mutation is found to be significantly preferentially expressed in patient 1 while TP53, PCCB and CCD6 mutations are found to be significantly preferentially expressed in patient 2. In addition, this analysis shows variants that are not expressed, variants that are expressed at equal levels as in the genome and variants where the germline variant is preferentially expressed (Fig 3). These results reveal that an analysis of somatic mutation is better complemented with allelic expression studies which show whether a mutated allele is expressed and the degree and direction of that expression. Mutated alleles that are not expressed likely have no impact on tumor progression.

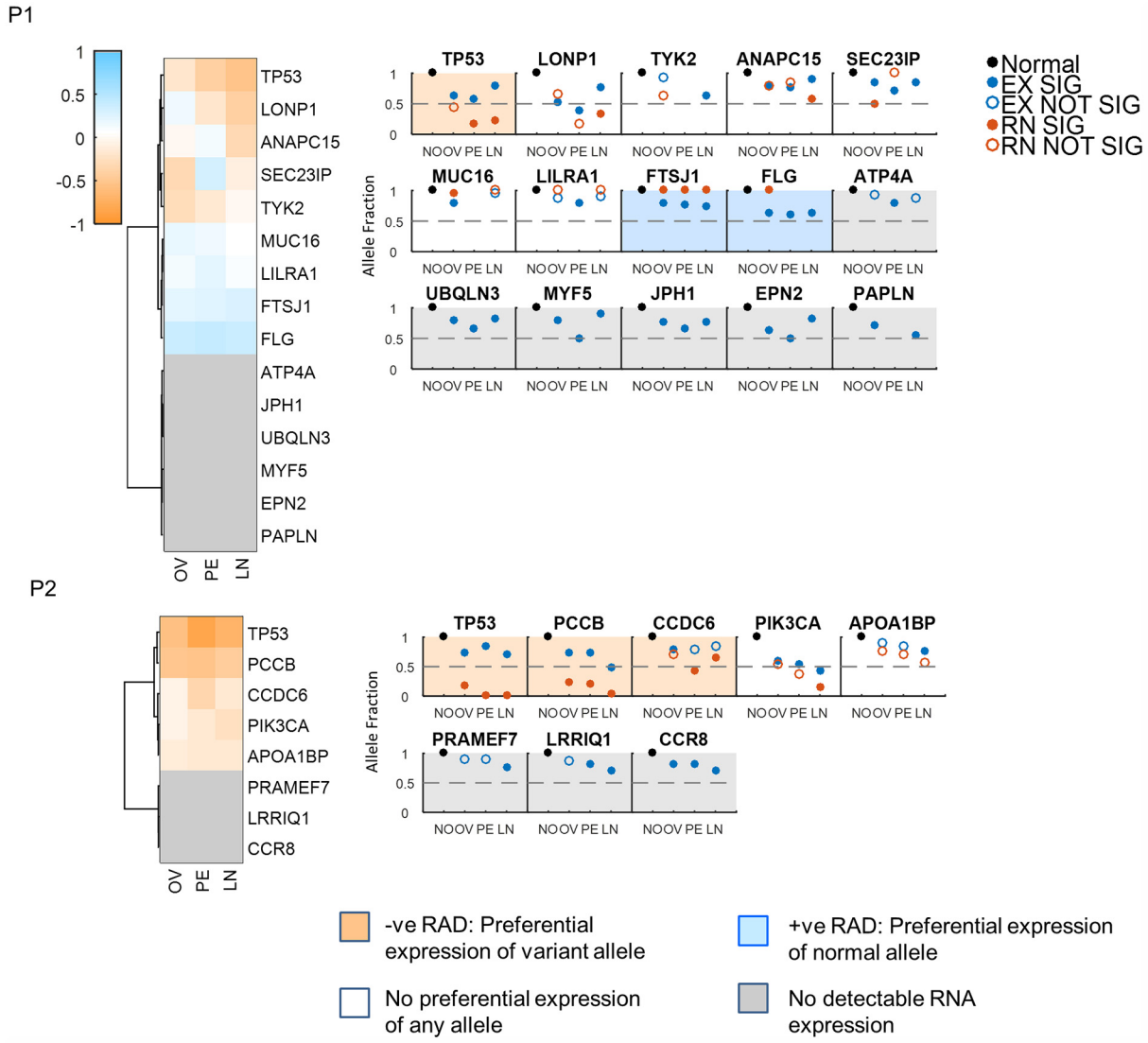


Fig 3. Somatic mutations preferential allele expression in patient 1 (P1) and patient 2 (P2). Clustering of the allele fraction differences for all somatic genes shows somatic mutations with preferential allele expression for the mutant allele (orange), no preferential expression (white), reference allele preferential expression (blue) and no expression of any allele (gray). Subplots for each patient show the exome (blue circles) and RNA (red circles) allele fraction for every site (NO is normal, OV is ovary, PE is peritoneum and LN is lymph node). Filled blue circles indicate significant difference from 1 and filled red circles indicate significant difference from the exome allele fraction for the given sites. Open circles indicate no significant difference. An absence of a circle indicates no data for that site. Plots are sorted from the highest allele fraction difference to the lowest. Genes with no RNA expression were arbitrarily assigned a highly negative allele fraction difference. Shaded plots indicate if there's significant preferential allele expression and the direction of the expression as stated in the legend. Note that shading was done as an aggregate of all sites within a patient. For example, in P2, one site in PIK3CA has a significant allelic difference (the LN) while there is no difference at other sites.

doi:10.1371/journal.pgen.1005755.g003

In addition to analyzing somatic mutations, we analyzed all variant alleles that consist of germline and somatic variants. For most variants, there is no significant difference in the exome and RNA allele fractions (Fig 4A). However, for a relatively small number of variants, there are significant allele fraction differences (Fig 4A and 4B). These variants are both germline and as previously stated somatic with the vast majority (~96%) being germline. Furthermore, using the normal data for two patients we estimated how many reference/alternate alleles would have been incorrectly assigned to be able to estimate the error for the third patient

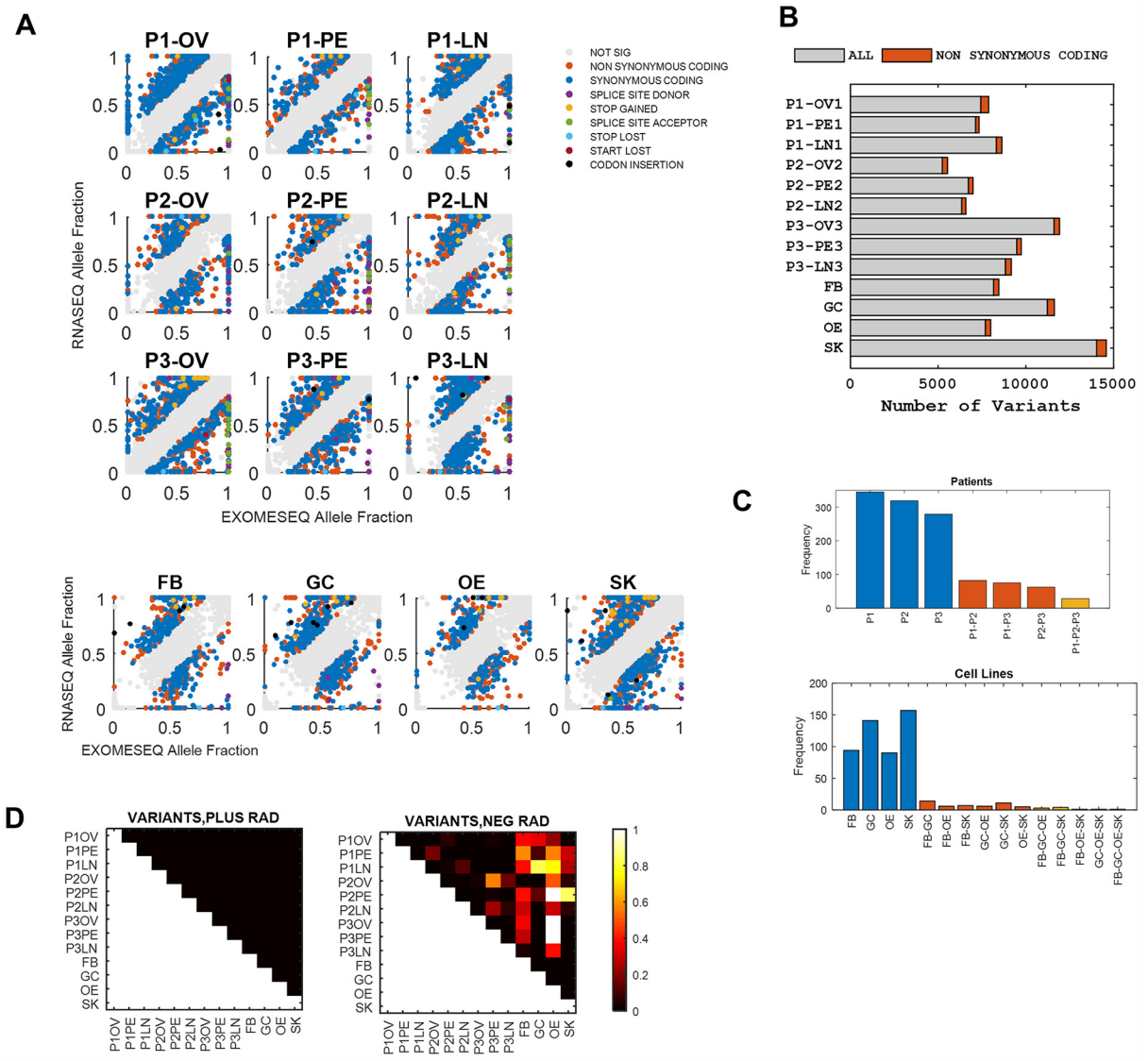


Fig 4. All variant preferential allele expression analysis. A) Scatter plots show the exome and RNA allele fraction for every patient and site in our study and for cells lines (FB, GC, OE, SK). Grey circles are variants with non-significant allelic differences while significant differences are indicated in colored circles as stated in the Legend. Note we imposed a cutoff of allelic difference of 0.2. B) The number of significant non-synonymous coding variants relative to all variants for patients and cell lines. C) The shared variants in patients and cell lines are shown as bar plots. D) Analysis of the significance of shared variants. The approximate p-value for the pair is shown in the heatmap with p-values of 0 being black and p-values of 1 being white. We analyzed the positive RAD variants and the negative RAD variants and found large difference between the cell line data and the patient site.

doi:10.1371/journal.pgen.1005755.g004

and the cell line data. As S2 Fig shows, the error ranges from 4.5%-9.4%. This means that patient 3 and cell line data likely have up to 10% of their alleles incorrectly called as alternate or reference. This would affect the direction of preferential allele expression in our analysis. However, for our global analysis this error would not significantly affect our major conclusions.

We also performed exome and RNA sequencing on four different cell lines (SKOV3, GOC2, fibroblasts and primary ovarian epithelial cells). As in the cell lines, we identify significant allele fraction differences in a small subset of variants (Fig 4A and 4B). Interestingly, while the same types of variants is observed in both cell lines and patient data, the cell lines show an enrichment in STOP gained variants compared to patient data (S3 Fig).

Shared Preferentially Expressed Alleles

To analyze the significance of the observed preferentially expressed variants, we compared the cell lines and patient shared genes. We observe a striking difference between cell lines and patient shared genes as the variant genes identified are very different between the two. The patient shared variants are substantial with 28 shared genes while only few genes are shared among the four different cell lines (Fig 4C). No gene is shared between cell lines and patients.

However, the number of significant variants is larger overall in patients which would lead to higher number of shared variants. To determine if the relatively high number of shared variants in patient data is due to chance, we performed a significance analysis. We selected variants randomly from the total pool of variants in the data and counted how many times these variants are shared between patients and cells. We then calculate a p-value corresponding to the number of times the shared random replicates overlaps with the observed shared replicates. As the p-value heatmap in Fig 4D shows, there is significant shared variants among patients and cell lines for the reference variants while there is significant shared variants for the alternate variants only among the patient data. This indicates that the set of shared alternate variants are significant in the context of patient tumors while the set of reference variants are significant for cell lines and patients. We therefore conclude that the set of preferentially expressed alternate alleles are specifically relevant to patient cancer data and focusing on that set could yield to insights into patient tumor biology.

To further understand if the alleles in our data are selected for we carried out a synonymous to nonsynonymous analysis for every patient and site. In this analysis, for every gene with at least 2 variants, we calculate the ratio of the number of non-synonymous to synonymous variants for the positive RAD, negative RAD and no-RAD variants. As shown in S4 Fig, in every patient and site, variants with positive and negative RAD generally have a non-syn/syn ratio greater or equal to the variants with no RAD. This is especially visible at the P1-LN site. This suggests that there could be selection for some of the positive and negative RAD variants. However, this type of analysis is limited in that there were few genes with enough variants to calculate synonymous and non-synonymous ratios as there is little variation within genes from the same patient. For most of the genes, selection cannot be identified using this method.

We next looked into the biological significance of the shared patient's alternate alleles. There were 28 genes shared across three patients (Figs 4C and 5A). Among the 28 shared genes, the majority are known to be involved in tumor biology based on IPA annotation of the genes and manual inspection. These include *TP53* (the highly mutated in epithelial ovarian cancer tumor suppressor), *MUC16* (also known as CA125) [25–27], *MKI67* [28–31], *LAMC2* [32–34] and ERBB2 [35]. In addition, the corresponding proteins localize to the extracellular space, plasma membrane, cytoplasm and nucleus.

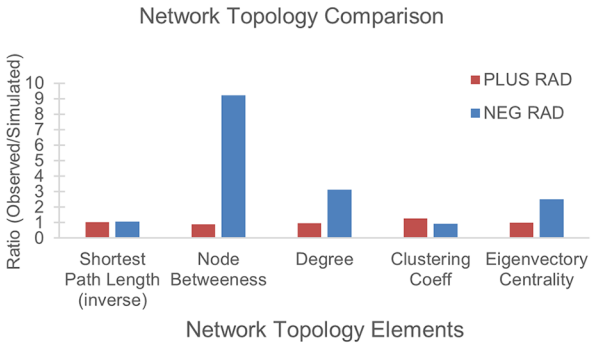
Two genes, *MUC16* and *MKI67*, show a large number of variants (38 and 9 different variants respectively). The large number of variants in these genes could be due to the size of these genes, alignment errors in repeat regions (though we filtered out multi-mapping reads and low quality reads) or naturally polymorphic positions or may be due to heterogeneity in different cells. We include them in the list of significant genes due to their known role in tumor biology and because some of the variants we identified have also been reported as somatic mutations in COSMIC for all two genes (Fig 5A). *MUC16* is known to be expressed in most serous ovarian carcinomas and may function like *MUC1* and *MUC4* in tumor cell growth, motility and tumorogenicity [26]. *MKI67* (Ki-67) is a well-known proliferation marker associating with prognosis [29, 36] and has been used as a target in an ovarian cancer model system [30].

We then performed a topological network analysis using TopoGA [37] and Jepetto [38] on the set of shared 28 genes with alternate allele preferential expression (negative RAD) and the

A

#	Genes	Average Patient Allele Fraction Difference			Gene Level Summary				Mutation Type Summary					DBSNP STATUS	
		P1	P2	P3	Mutation Types	Unique Genome Locations	Unique RS ids	Number in Cosmic	NON SYNONYMOUS CODING	SPLICE SITE ACCEPTOR	SPLICE SITE DONOR	START LOST	STOP GAINED	NON DBSNP	DBSNP
1	ATXN3				1	3	3	2	●					1	2
2	AZIN1				1	1	1	0	●					1	0
3	BDP1				1	5	5	5	●					0	5
4	CDK11A				1	1	1	1	●					0	1
5	CDK11B				2	3	3	3	●		●			0	3
6	ERBB2				2	3	3	0	●		●			1	2
7	GNAS				1	1	1	0		●				0	1
8	GPX7				1	3	1	0	●					3	0
9	HIVEP1				1	3	3	1	●					1	2
10	IFT74				1	2	2	0	●					0	2
11	IL1R1				1	1	1	1	●					0	1
12	KMT2C				2	3	3	2	●				●	1	2
13	LAMC2				1	2	2	0	●					0	2
14	MKI67				1	9	9	8	●					0	9
15	MMP7				2	2	2	1	●		●			1	1
16	MRPL35				1	1	1	1	●					0	1
17	MUC16				1	38	33	21	●					6	32
18	NACA				1	1	1	0			●			1	0
19	PABPC3				1	1	1	1	●					0	1
20	PRSS3				1	4	4	4	●					1	3
21	RARRES3				1	1	1	0	●					1	0
22	SHROOM3				1	1	1	0	●					0	1
23	TMTC4				1	1	1	1	●					0	1
24	TNC				1	4	4	4	●					0	4
25	TP53				2	3	1	3	●		●			3	0
26	TUBA1B				1	1	1	1			●			0	1
27	UQCRC2				1	1	1	0		●				1	0
28	ZP3				1	1	1	1	●					0	1

B



C

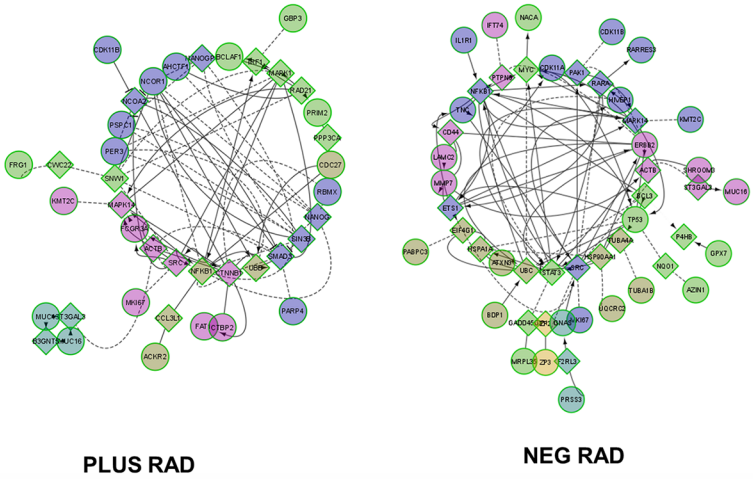


Fig 5. Network analysis of shared patient plus and negative RAD sets. A) List of 28 genes with patient and gene level summaries of the different variants. B) Toplogical analysis of network associations based on known biology among the set of positive and negative RAD genes. C) Association between genes of the negative and positive RAD sets. Each symbol is a gene with colors corresponding to the clustering analysis within that gene. Circle symbols indicate genes shared within our dataset while squares indicate associated genes not in our dataset. Solid lines indicate direct interactions while dotted lines indicate indirect interactions. Note the higher frequency of direct interactions among negative RAD genes.

doi:10.1371/journal.pgen.1005755.g005

set of shared 26 genes with reference allele preferential expression (positive RAD). As Fig 5B shows, there is a clear increase in connectedness among the alternate allele preferentially expressed genes. Fig 5C also shows a plot of the known genetic interactions from the reactome functional interaction database [39, 40] and there are many more direct interactions for the alternate allele preferentially expressed genes than for the reference allele preferentially expressed genes.

Overall, our results strongly indicate that, in contrast to the set of preferentially expressed reference alleles, the shared set of preferentially expressed alternate/variant alleles play a role in ovarian cancer biology since the majority of the identified genes are involved in known cancer processes.

Additionally, we find that the expression level for most of these 28 genes is remarkably consistent across different patients and sites (S5 Fig). This indicates that these genes with significant allelic bias cannot be identified from gene expression studies.

Patient and Site Specific Preferentially Expressed Alleles

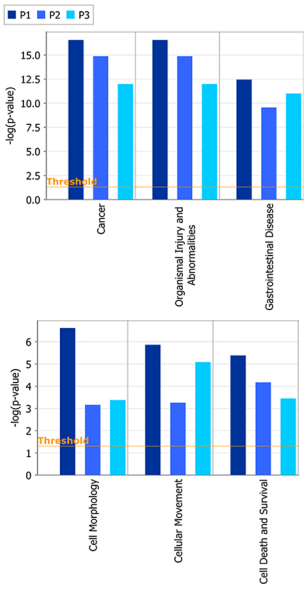
We then analyzed the genes with preferentially expressed alleles for each patient individually using Ingenuity pathway analysis. Each patient's significant genes are associated with similar cancer related diseases and processes (Fig 6A). For example for the Cancer category, for patients 1, 2 and 3 at least half of the genes were in cancer associated genes. Note that the gastrointestinal disease reported by IPA include cancer diseases. Similarly there is enrichment for functions such as DNA repair and cell growth for every patient. Our analysis therefore identifies many genes that are known to be involved in tumor development. This is the case whether looking at shared genes or whether looking at private genes.

The identification of a large number of cancer-related genes for each patient also allows us to ask whether any of these genes are targetable with known drugs. We are able to identify for each patient several drugs that directly target genes (Fig 6B). Many of these drugs are or have been in clinical trials or are widely used for current cancer therapy as indicated in Fig 6B in bold. Notably, preferential allele expression in aurora kinase genes (AURKA) is significant in two patients. AURKA are targets of new drugs currently undergoing clinical trials [41].

Looking further at networks of drug targetable genes (Fig 6C) we observe that significant genes in our analysis form an interconnected network and that drugs exist that can target different points in this network. Targeting these genes, which form central networks in cells, are likely to result in measurable effects on cancer cells.

Finally, since our data includes primary and metastatic samples, we also analyzed the allele fraction differences at a site-specific level to identify any site-specific patterns. While there was little overlap between genes at the site level (Fig 7A), the disease pathways and biological functions were found to be the same in the different sites (Fig 7B). We then performed a hierarchical clustering analysis on significant allele fraction differences for all patients and sites (Fig 7C). As expected, patients clustered together but every sample and site clustered independently at the variant level. We also identified shared patterns across all possible clusters of negative allele fraction differences. These clusters represent sets of variants that show similar patterns of significant differential allele expression (Fig 7D). These clusters are the OV, PE, LN, OVPE, OVLN, PELN and OVPELN clusters. The PE, LN and PELN clusters are especially interesting as they are metastatic clusters. Interestingly, *TP53* appears in the PELN cluster indicating that metastasis samples (PE and LN) are overexpressing the variant TP53 more than in the ovary sample.

A



B

Genes	P1	P2	P3	Drug
MMP7				marimastat
TP53				**CGM097, kevetrin**
C3				IgG
IL1R1				anakinra
PRIM2				cytarabine, others
ERBB2				**trastuzumab, others**
HSP90AB1				**alvespimycin, retaspimycin**
IL7R				recombinant human interleukin-7
FCGR3B				IgG
FDFT1				lapaquistat
MSLN				**amatuximab**
AURKA				**several drugs**
ATIC				gemcitabine
MMP9				**GS-5745**
ANPEP				**CNGRC peptide-TNF alpha conjugate**
GABRE				pagoclone, others
COL6A3				collagenase clostridium histolyticum
COL6A2				collagenase clostridium histolyticum
KDR				AEE 788, others
CFTR				crofelemer, ivacaftor
PDE5A				dyphylline, others
MUT				hydroxocobalamin, cyanocobalamin
PSMB10				**carfilzomib, others**
TSPO				CGS-8216, others
KCNMB1				**tedisamil**
COL6A1				collagenase clostridium histolyticum
COL4A1				collagenase clostridium histolyticum
PML				**arsenic trioxide**
IL1B				**canakinumab, others**
TPMT				collagenase clostridium histolyticum
COL12A1				collagenase clostridium histolyticum
SLC25A5				clodronic acid
PIK3CA				**SF 1126, others**
PSEN1				tarenflurbil
CD52				**alemtuzumab, others**
MMP14				marimastat
IMPDH2				**thioguanine, others**
CAT				fomepizole
PDGFRB				**nilotinib, others**
TLR4				resatorvid
ALAD				delta-aminolevulinic acid

C

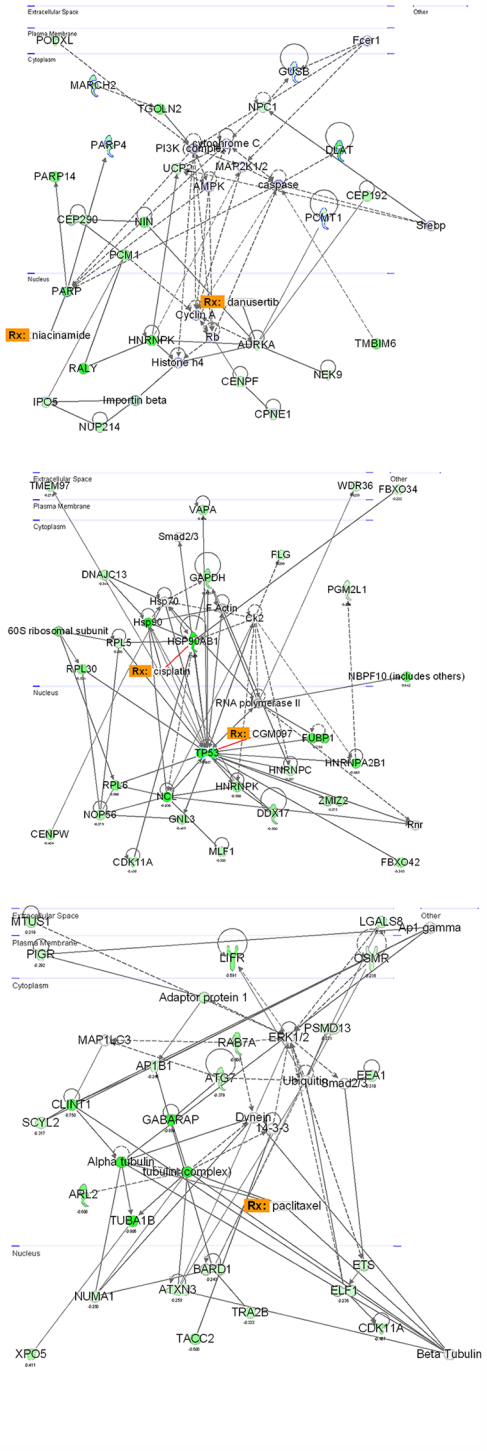


Fig 6. Patient specific analysis of negative RAD genes. A) Top biological pathways molecular functions for every patient's set of significant preferentially expressed alleles. B) Druggable targets in every patient. The bar plots show the extent of preferential allele expression and the list is sorted by the average of allelic expression from highest to lowest. Drugs in bold type are used in current anticancer treatment or undergoing clinical anticancer trials. C) Network plots of a selected pathway in every patient showing a druggable target.

doi:10.1371/journal.pgen.1005755.g006

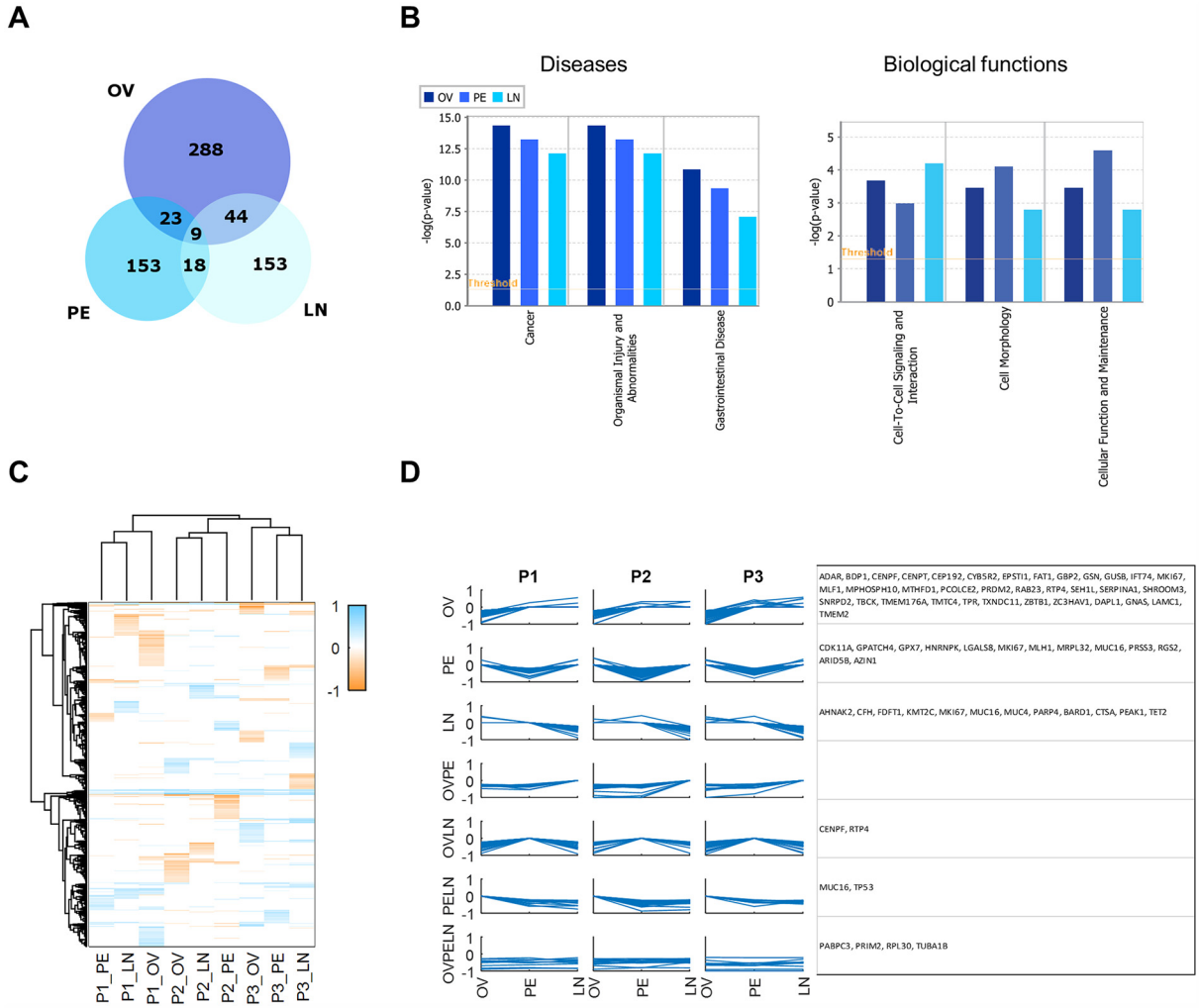


Fig 7. Site-specific preferential allele expression. A) Overlap of preferentially expressed variant alleles for the different sampled sites (OV, PE, LN). Numbers in intersecting or non-intersecting regions indicate the number of shared genes. B) Ingenuity pathway analysis of the different genes in each site showing the top 3 predicted diseases and the biological functions for each site. Note that gastrointestinal disease indicates cancer of the gastrointestinal system. Threshold (orange line) is set at a p-value of 0.05. C) Hierarchical clustering analysis of RNA-exome allele fraction differences expression differences show that different sites have largely different allelic expression patterns. As in previous figures, a positive allelic difference indicates a reference allele is preferentially expressed while a negative allelic difference indicates an alternate allele is preferentially expressed. Patients are also different with few shared alleles. D) Parallel plots of allele fraction differences for different patients and different site expression patterns. The site or combination of sites where there is a negative allele fraction difference is indicated by the site names (OV, PE, LN) for ovary, peritoneum and lymph node. Gene names that appear in at least two patients for a given pattern are indicated to the right of the parallel plot.

doi:10.1371/journal.pgen.1005755.g007

Discussion

The keys results of our analysis are the identification of a large number of shared genes showing preferential variant allele expression among three different patients. We identify 28 shared genes many of which are known to be involved in ovarian cancer or other cancers and occur in cellular pathways that are relevant to tumor biology. These shared elements could form the basis for the biological similarities among ovarian cancers and lead to new therapeutic targets and better patient targeting of current therapies.

The observed preferential allele expression could indicate potential selection for that allele due to providing growth or survival benefits to the tumor. Alternative causes for significant

preferential allele expression include artifacts of library preparation, alignment or sequencing. Preferential allele expression may also occur without biological or tumor significance. However, the evidence indicates that our analysis is capturing a biologically relevant set of genes. First, the number of variants that we identify as being significant is a small proportion of the total variants (Fig 4A and 4B) which indicates that there is no large systematic issue. Second, we validated the exome read ratios with SNP6 array data (where possible) and find excellent agreement between the array and exome data (S6 Fig). Third, we identify a large number of shared genes between different patients. Within these shared genes there are many variant types indicating that there is no bias for specific variant types. Some variants are the same in all patients while some variants are different in two or three other patients. Fourth, many of the genes we identify have very well known or plausible effects on tumor biology. Fifth, as discussed previously and below, many of the genes we identify with significant allele fraction differences are known drug targets. Sixth, comparing cell line data to patient data reveals large differences in the type of variants observed. There is also overall large similarity within patients than within cell lines suggesting that the preferentially expressed variant alleles in patients are specific. We therefore conclude that while there could be some biases due to low coverage or a higher error rate in one dataset, the majority of genes we identify are relevant to tumor biology and merit further investigation.

There are multiple mechanisms that can give rise to preferential allele expression. Copy number changes, including loss of an allele (LOH, loss of heterozygosity), RNA editing and allele specific methylation are three such mechanisms. However, our analysis methods that consider both the genomic and transcriptomic content would likely not detect preferential allele expression solely caused by copy number changes as we subtract the RNA read fraction from the DNA read fraction. For RNA editing, while our analysis would detect novel RNA only alleles, we restrict in this work our analysis to alleles that occur in both the exome and RNA sequencing data. This is because additional evidence would be needed to be have confidence in the novel RNA alleles. Allele specific methylation [42] could explain the preferential allele expression in our data as one allele could be methylated therefore reducing the expression of that allele relative to the unmethylated allele. Future experiments combining genome, transcriptome and methylation sequencing could determine the number and identity of preferentially expressed alleles modulated by methylation.

Our filtering procedure was stringent but it is possible that some of our data is due to different types of technical errors. Improving this analysis further could involve excluding reads from error-prone regions, applying more robust statistical methods and validating the expression data with site-specific qPCR. In addition, while we focused in this study on non-synonymous mutations or other high impact mutations, analyzing the synonymous mutations could lead to additional insight into tumor biology as synonymous mutations have been shown recently to act as drivers in tumors [43].

Another interesting subset of genes in our data are genes affecting cell adhesion and migration. The set of 28 shared and significant genes includes *LAMC2* [33, 34], MMP7 [44] and TNC [45]. We have previously described a large enrichment in somatic mutations in adhesion genes using the TCGA ovarian cancer data [46] while others have identified cell adhesion gene enrichment in a multidimensional study of ovarian cancer data [47]. Cell adhesion and migration are essential processes in tumor development and metastasis and it would be interesting to further investigate the functional significance of the variants we identify.

One notable advantage of applying CGTS to tumor samples is to be able to identify significant preferentially expressed germline variants. Somatic mutation analysis and standard gene expression studies cannot investigate germline variation. Germline variants are the vast majority of variants in cells and understanding their role in tumor biology could be critical to

designing effective therapies. Another advantage of CGTS is that there is no requirement for large numbers of patients—it is possible to identify significant preferential allele expression in one patient.

Aside from potentially uncovering cancer drivers and suggesting patient-specific therapies, the observation that there is significant overlap between patients in the genes that have preferentially expressed alleles suggests that there is less heterogeneity at this level than seen with somatic mutation analysis. This increased homogeneity at the allelic expression level could impact the clustering of patients and has therapeutic implications as more patients can be treated with the same drug.

Overall, site analysis seems to indicate that different sites while having different genes utilize similar pathways. Perhaps the differences between primary and metastasis sites are due to gene expression changes or post-translation modifications. These results, while puzzling, are quite interesting. Indeed, the previous two studies on multiple site sampling [4, 5] demonstrated a high degree of mutational heterogeneity but most patients have a good initial response at all sites to chemotherapy potentially indicating shared biological features despite mutational heterogeneity.

Understanding tumor biology to impact clinical care requires the integration of multiple datasets in clinically meaningful ways. This work adds an additional novel dimension to the large body of work in tumor biology by analyzing preferential allele expression using a combined genome and transcriptome sequencing approach. The striking result of finding a set of shared biologically relevant genes with preferentially expressed variant alleles suggests that these genes may be driver genes. Future work characterizing these genetic variants individually and determining mechanisms of preferential allele expression would lead to greater understanding of the functional significance of these variants in tumor biology.

Methods

Ethics Statement

All samples were collected at the department of Gynecologic Oncology at the Institut Claudius Regaud. The project was reviewed and approved by the Institut Claudius Regaud Human research Ethics Committee. All patients included in the study gave informed written consent prior to surgery.

Sample Collection

Three patients with Stage 3C high grade serous ovarian cancer were recruited for this study before any treatment at the Institut Claudius Regaud, Toulouse. All displayed metastasis in the lymph node and peritoneum. During primary cytoreductive surgery, tissue samples from the ovary, peritoneum and lymph node were collected and snap frozen. Biopsies were macrodissected and snap frozen sections were controlled and samples with 80% of tumor cells were selected. DNA and RNA was extracted with Qiagen RNA/DNA extraction kit. Additional clinical details are in S1 Text. Patients and sites are abbreviated (P1, P2, P3 for patients 1, 2, 3 and OV, PE, LN for ovary, peritoneum and lymph node).

Ovarian cancer cell line SKOV3 (HTB-77) was purchased from ATCC, primary ovarian cancer cell line GOC2 propagated in-house and human fibroblasts were maintained in culture (DMEM high glucose [Hyclone, Thermo Scientific], 10% FBS [Hyclone, Thermo Scientific], 1% Penicillin-Streptomycin-Amphotericyn B solution [Sigma], 1X Non Essential Amino-Acid [Hyclone, Thermo Scientific]). Human primary ovarian epithelial cells from ScienCell were cultured in poly-L-lysine-coated culture vessel (2 μg/cm2, T-75 flask) following ScienCell recommendations (Ovarian Epithelial Cell Medium (OEpiCM, Cat. No. 7311), 1% Ovarian Epithelial Cell Growth Supplement (OEpiCGS, Cat. No.7352), 1% penicillin/streptomycin

solution (p/s, Cat.No 0503). Cultures were incubated in humidified 5% CO2 incubators and the media was replaced every 3 days. RNA and DNA were isolated using Qiagen Allprep DNA/RNA miniprep kit Cat. No. 80204 following manufacturer instructions and stored at -80 degrees Celsius before sequencing. Cells lines are abbreviated SK (SKOV3), GC (GOC2), FB (Fibroblast) and OE (ovarian epithelial).

Sequencing and Alignment

RNA and exome sequencing was performed at Weill Cornell Medical College—Qatar, Genomics core. Exome capture was done using Agilent's 38 mB SureSelect Human All Exon kit (patient tumor samples) and SureSelectXT2 Human All Exon V5 (cell lines and normal patient samples). Paired end, 100bp sequencing was done on an Illumina Genome Analyzer IIx (patient tumor samples) and Illumina HiSeq 2500 (cell lines and normal patient samples). Reads were aligned using BWA (version 0.7.9a) [48], indexed with samtools (version 0.1.18) [49], processed with Picard mark duplicates (http://picard.sourceforge.net, version 1.110) and realigned with GATK (version 3.1.1) [50, 51]. GATK bundle 2.8 was used for the reference genome hg19 and the realignment data. RNA was processed with Nugen Ovation v2 kit (patient tumor samples) and Nugen Ovation Single Cell RNA-Seq System (cell lines and normal patient samples) and 100bp sequencing was done on an Illumina HiSeq 2000 (patient tumor samples) and Illumina HiSeq 2500 (cell lines and normal patient samples). Resulting reads were aligned using RNA Star (version 2.4.0g1) in 2-pass mode [52] to the reference genome hg19 (S1 Fig) and deduplicated with Picard mark duplicates (version 1.110). Single nucleotide variants with mapping quality >20 and base quality > 20 were called using VARSCAN (version 2.3.7) [53, 54] from samtools mpileup output [55] combining all deduplicated and filtered exome and RNA samples and then annotated with SNPEFF (version 3.6) to obtain a VCF. Alignment and post processing read counts are shown in S1 Fig. The annotated VCF was then analyzed further with custom scripts.

Allele Fraction Differences

Allele fractions were calculated for single nucleotide variants identified in exome and RNA sequencing data. We limited our analysis to reads with base and mapping quality greater than or equal to 20 and non-duplicated reads for both exome and RNA sequencing data. We further focused only on protein coding variants. Allele fractions are defined as the ratio of the number of reads of the reference allele to the total number of reads at a site. The Fisher Exact Test with the mid-P correction was used to obtain p-values between the RNA and Exome data (significance was set at a p-value less than 0.1). In addition, we imposed a filter that there should be at least eight reads for both the exome and RNA sequencing data to account for errors in sequencing and alignment [56]. The allele fraction difference (RAD) is calculated as the RNA allele fraction minus the exome allele fraction. Allele fraction differences range from -1 to 1 with negative differences meaning that the variant allele is preferentially expressed and positive differences meaning that the reference allele is preferentially expressed. Significant alleles refer to those alleles with significant negative preferential allele expression. Significant genes are those which have at least one variant allele has statistically significant preferential allele expression. Where normal sample (non-tumor) data is not available, we consider an allele to be likely germline if it also exists in dbSNP.

Variant Annotation

SNPeff [57] was used to annotate the variants to identify protein coding variants. We use the same 'IMPACT' categorization as SNPEFF (High, Moderate, Low and Modifier). We used

Ingenuity Pathway Analysis software (Ingenuity Systems) to annotate and functionally characterize gene lists and to identify drug targets. To identify SNPs that were in COSMIC [58], we downloaded the COSMIC complete export list (v72) and cross-referenced it with our variants based on genomic coordinates.

Shared Gene Analysis

Identifying shared genes between groups is important as shared genes could point to common mechanisms. However, in sets where the number of elements is limited such as gene sets, it is possible to have shared genes that are likely to occur due to chance. To determine in our data if the number of shared genes or variants are significant, we calculated the number of shared variants from 1000 random selections of variants in our variant data (Shared Random) and compared that to the number of observed shared variants (Shared Observed). We then calculate a pseudo p-value which is the number of times the shared random selection exceeds or equals the shared observed number divided by the total number of random trials. A value of 0 means that in no random selection trial was the shared number of genes greater than the observed number of shared genes. A value of 1 means that in every random selection trial the number of shared genes was greater than or equal to the observed shared gene number. Custom Matlab code was developed for this analysis.

Network Toplogy and Reactome Functional Interactions

Shared genes in both the patient positive and negative RAD sets were input into Jepetto [38], a Cytoscape interface for TopoGA [37] and topology calculated using the large dataset settings. Functional interactions were determined by using the Reactome Functional Interactions [39, 40] Cytoscape plugin using the 2014 dataset and using linker genes.

Non-synonymous Variant Enrichment Analysis

This analysis counted at a gene level the frequency of non-synonymous and synonymous variants for different RAD sets: positive RAD, negative RAD and no RAD. For this analysis to be possible, genes must have more than one variant. Data was analyzed at numbers of variants greater than or equal to 2 and greater than or equal to 4. Custom Matlab code was used for this analysis.

Gene Expression

RNA sequencing data was used to estimate relative gene expression. Alignments (see sequencing) were processed with FeatureCounts of Rsubread (version 1.18) [59] in paired-end mode, excluding overlapping reads and overlapping genes to map reads to genes (gencode v19 gene models). Filtration (remove all genes which have less than 0.1 counts per million) and normalization was done with edgeR (version 3.4.2) [60] to obtain reads per kilobase per million (rpkm) values for each gene. Patient and cell line data was analyzed identically.

Visualization

GenomeBrowse (Golden Helix) was used to visualize the TP53 gene region. Cytoscape (version 3.2.1) [61] was used to visualize networks in Fig 5. Network diagrams in Fig 7 were made using Ingenuity software (Qiagen). Proportional Venn diagrams were made using BioVenn [62] and adjusted with Inkscape (version 0.91) Hierarchical clustering was done with built-in Matlab (version R2015a) (Mathworks) commands. Pileups, gene subplots, gene expression plots and parallel plots were made with custom Matlab scripts.

Supporting Information

S1 Fig. Exome and RNA sequencing read counts.
(TIF)

S2 Fig. Shows what are the significant changes when considering the control corrected differential allele ratios. The open colored circles are those significant differential allele expression before expression and the filled circles are the same after expression. Bottom table shows the numbers of the differences and these range from 4.5% to 9.4%.
(TIF)

S3 Fig. Allele fraction difference histograms for different types of mutations for A) Patients and B) Cell lines. Patient, sites and cell line names are indicated on the left. The allele fraction difference on the x-axis ranges from -1 (only variant allele is expressed) to +1 (only normal allele is expressed). The frequency is indicated on the y-axis. Allele fraction differences less than -0.2 or greater than 0.2 were removed. Panels with horizontal lines have no variant of that type.
(TIF)

S4 Fig. Cumulative density function plots of non-synonymous over synonymous ratios (non-syn/syn) for genes that have A) at least 2 significant variants or B) at least 4 significant variants. The grey, orange and blue lines are as indicated in the legend. Two notable observations are 1) no preferential allele expression always has low non-syn/syn ratios relative to at least one form of preferential allele expression and 2) there is a marked increase in non-syn/syn ratios in P1LN sample and 3) there are higher non-syn/syn ratios for +RAD genes.
(TIF)

S5 Fig. Patient and cell line gene expression levels for the shared 28 genes. The x-axis of each panel shows the different sites (OV, PE, LN), the line colors correspond to different patients as indicated in the legend, and in letters (F, G, O, S) indicate the gene expression value for the cell line samples as indicated in the legend. The y-axis is the log2 (rpkm). Note the relative similarity of gene expression for most patients and sites. Exceptions to this similarity are ERRB2 which is high in patient 2 and MMP7 which shows marked changes in different sites.
(TIF)

S6 Fig. Validation of Exome allele fractions with SNP6 data. Note that the exome allele fraction is based on the reference allele (as used in the manuscript) while the SNP6 frequency is based on the minor allele (as is customary in SNP array analysis). The red points are those where the minor/major allele needed to be switched as the minor allele matched the reference allele and the major allele matched the alternate allele. The green and black points represent sites where at least one of the RNA sequencing alleles did not match any of the SNP array alleles. The negative correlation is because we are comparing Exome Reference Allele Frequency to SNP array Minor Allele Frequency.
(TIF)

S1 Text. Clinical description of patients.
(DOCX)

Acknowledgments

We would like to thank members of the Rafii and Malek labs and the WCMC-Q bioinformatics group members for many helpful discussions.

Author Contributions

Conceived and designed the experiments: NMH DQ AR AM. Performed the experiments: AM EM LP GF DQ JAM AR HAF PP IAA EAD YAM. Analyzed the data: NMH AM HGK CM JAM AR. Contributed reagents/materials/analysis tools: JAM. Wrote the paper: NMH AM AR HAF.

References

1. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314(5797):268–74. doi: 10.1126/science.1133427 PMID: 16959974
2. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502(7471):333–9. Epub 2013/10/18. doi: 10.1038/nature12634 PMID: 24132290; PubMed Central PMCID: PMC3927368.
3. Network TCGAR. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474(7353):609–15. doi: 10.1038/nature10166 PMID: 21720365
4. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. J Pathol. 2013; 231(1):21–34. doi: 10.1002/path.4230 PMID: 23780408; PubMed Central PMCID: PMC3864404.
5. Hoogstraat M, de Pagter MS, Cirkel GA, van Roosmalen MJ, Harkins TT, Duran K, et al. Genomic and transcriptomic plasticity in treatment-naive ovarian cancer. Genome Res. 2014; 24(2):200–11. Epub 2013/11/14. doi: 10.1101/gr.161026.113 PMID: 24221193; PubMed Central PMCID: PMC3912411.
6. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res. 2008; 14(16):5198–208. doi: 10.1158/1078-0432.ccr-08-0196 PMID: 18698038
7. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic Variation in Human Gene Expression. Science. 2002; 297(5584):1143-. doi: 10.1126/science.1072545 PMID: 12183620
8. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, et al. Allelic variation in gene expression is common in the human genome. Genome Res. 2003; 13(8):1855–62. PMID: 12902379
9. Smith RM, Webb A, Papp AC, Newman LC, Handelman SK, Suhy A, et al. Whole transcriptome RNA-Seq allelic expression in human brain. BMC Genomics. 2013; 14(1). doi: 10.1186/1471-2164-14-571
10. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009; 25(24):3207–12. doi: 10.1093/bioinformatics/btp579 PMID: 19808877
11. Crum C, Drapkin R, Miron A, Ince T, Muto M, Kindelberger D, et al. The distal fallopian tube: a new model for pelvic serous carcinogenesis. [Miscellaneous Article]. Current Opinion in Obstetrics \& Gynecology February 2007. 2007; 19(1).
12. Kim J, Coffey D, Creighton C, Yu Z, Hawkins S, Matzuk M. High-grade serous ovarian cancer arises from fallopian tube in a mouse model. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(10). doi: 10.1073/pnas.1117135109
13. Vang R, Shih I-M, Kurman R. Fallopian tube precursors of ovarian low- and high-grade serous neoplasms. Histopathology. 2013; 62(1). doi: 10.1111/his.12046
14. Perets R, Wyant G, Muto K, Bijron J, Poole B, Chin K, et al. Transformation of the fallopian tube secretory epithelium leads to high-grade serous ovarian cancer in Brca;Tp53;Pten models. Cancer Cell. 2013; 24(6). doi: 10.1016/j.ccr.2013.10.013
15. Yang-Hartwich Y, Gurrea-Soteras M, Sumi N, Joo W, Holmberg J, Craveiro V, et al. Ovulation and extra-ovarian origin of ovarian cancer. Scientific Reports. 2014; 4. doi: 10.1038/srep06116
16. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA: a cancer journal for clinicians. 2009; 59(4):225–49. Epub 2009/05/29. doi: 10.3322/caac.20006 PMID: 19474385.
17. Benedetti-Panici P, Greggi S, Maneschi F, Scambia G, Amoroso M, Rabitti C, et al. Anatomical and pathological study of retroperitoneal nodes in epithelial ovarian cancer. Gynecol Oncol. 1993; 51(2):150–4. Epub 1993/11/01. doi: 10.1006/gyno.1993.1263 PMID: 8276287.
18. Morice P, Joulie F, Camatte S, Atallah D, Rouzier R, Pautier P, et al. Lymph node involvement in epithelial ovarian cancer: analysis of 276 pelvic and paraaortic lymphadenectomies and surgical implications. Journal of the American College of Surgeons. 2003; 197(2):198–205. Epub 2003/08/02. doi: 10.1016/S1072-7515(03)00234-5 PMID: 12892797.

19. Onda T, Yoshikawa H, Yokota H, Yasugi T, Taketani Y. Assessment of metastases to aortic and pelvic lymph nodes in epithelial ovarian carcinoma. A proposal for essential sites for lymph node biopsy. Cancer. 1996; 78(4):803–8. Epub 1996/08/15. doi: 10.1002/(SICI)1097-0142(19960815)78:4<803::AID-CNCR17>3.0.CO;2-Z PMID: 8756375.

20. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res. 2008; 68 (13):5478–86. Epub 2008/07/03. doi: 10.1158/0008-5472.CAN-07-6595 PMID: 18593951.

21. Malek JA, Martinez A, Mery E, Ferron G, Huang R, Raynaud C, et al. Gene expression analysis of matched ovarian primary tumors and peritoneal metastasis. Journal of translational medicine. 2012; 10:121. Epub 2012/06/13. doi: 10.1186/1479-5876-10-121 PMID: 22687175; PubMed Central PMCID: PMC3477065.

22. Malek JA, Mery E, Mahmoud YA, Al-Azwani EK, Roger L, Huang R, et al. Copy Number Variation Analysis of Matched Ovarian Primary Tumors and Peritoneal Metastasis. PLoS ONE. 2011; 6(12). doi: 10.1371/journal.pone.0028561

23. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013; 501 (7467):355–64. Epub 2013/09/21. doi: 10.1038/nature12627 PMID: 24048068.

24. Rose PG, Piver MS, Tsukada Y, Lau TS. Metastatic patterns in histologic variants of ovarian cancer. An autopsy study. Cancer. 1989; 64(7):1508–13. PMID: 2776109

25. Reinartz S, Failer S, Schuell T, Wagner U. CA125 (MUC16) gene silencing suppresses growth properties of ovarian and breast cancer cells. Eur J Cancer. 2012; 48(10):1558–69. doi: 10.1016/j.ejca.2011.07.004 PMID: 21852110

26. Thériault C, Pinard M, Comamala M, Migneault M, Beaudin J, Matte I, et al. MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis. Gynecol Oncol. 2011; 121(3):434–43. doi: 10.1016/j.ygyno.2011.02.020 PMID: 21421261

27. Chen S-H, Dallas MR, Balzer EM, Konstantopoulos K. Mucin 16 is a functional selectin ligand on pancreatic cancer cells. FASEB J. 2012; 26(3):1349–59. doi: 10.1096/fj.11-195669 PMID: 22159147

28. Pavelin S, Becic K, Forempoher G, Mrklic I, Pogorelic Z, Titlic M, et al. Expression of Ki-67 and p53 in meningiomas. Neoplasma. 2013; 60(5):480–5. doi: 10.4149/neo_2013_062 PMID: 23790165

29. Jin Q, Zhang W, Qiu X-g, Yan W, You G, Liu Y-w, et al. Gene expression profiling reveals Ki-67 associated proliferation signature in human glioblastoma. Chinese medical journal. 2011; 124(17):2584–8. PMID: 22040407

30. Rahmanzadeh R, Rai P, Celli JP, Rizvi I, Baron-Lühr B, Gerdes J, et al. Ki-67 as a molecular target for therapy in an in vitro three-dimensional model for ovarian cancer. Cancer Res. 2010; 70(22):9234–42. doi: 10.1158/0008-5472.can-10-1190 PMID: 21045152

31. Tawfik K, Kimler BF, Davis MK, Fan F, Tawfik O. Ki-67 expression in axillary lymph node metastases in breast cancer is prognostically significant. Human Pathology. 2013; 44(1):39–46. doi: 10.1016/j.humpath.2012.05.007 PMID: 22939959

32. Okuma E, Ohishi Y, Oda Y, Aishima S, Kurihara S, Nishimura I, et al. Cytoplasmic and stromal expression of laminin γ 2 chain correlates with infiltrative invasion in ovarian mucinous neoplasms of gastrointestinal type. Oncology reports. 2010; 24(6):1569–76. PMID: 21042753

33. Masuda R, Kijima H, Imamura N, Aruga N, Nakazato K, Oiwa K, et al. Laminin-5γ2 chain expression is associated with tumor cell invasiveness and prognosis of lung squamous cell carcinoma. Biomed Res. 2012; 33(5):309–17. PMID: 23124251

34. Katayama M, Sanzen N, Funakoshi A, Sekiguchi K. Laminin gamma2-chain fragment in the circulation: a prognostic indicator of epithelial tumor invasion. Cancer Res. 2003; 63(1):222–9. PMID: 12517801

35. Chmielecki J, Ross JS, Wang K, Frampton GM, Palmer GA, Ali SM, et al. Oncogenic alterations in ERBB2/HER2 represent potential therapeutic targets across tumors from diverse anatomic sites of origin. Oncologist. 2015; 20(1):7–12. doi: 10.1634/theoncologist.2014-0234 PMID: 25480824; PubMed Central PMCID: PMCPMC4294606.

36. Martin B, Paesmans M, Mascaux C, Berghmans T, Lothaire P, Meert AP, et al. Ki-67 expression and patients survival in lung cancer: systematic review of the literature with meta-analysis. Br J Cancer. 2004; 91(12):2018–25. doi: 10.1038/sj.bjc.6602233 PMID: 15545971

37. Glaab E, Baudot A, Krasnogor N, Valencia A. TopoGSA: network topological gene set analysis. Bioinformatics. 2010; 26(9):1271–2. doi: 10.1093/bioinformatics/btq131 PMID: 20335277; PubMed Central PMCID: PMC2859135.

38. Winterhalter C, Widera P, Krasnogor N. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks. Bioinformatics. 2014; 30(7):1029–30. doi: 10.1093/bioinformatics/btt732 PMID: 24363376; PubMed Central PMCID: PMC3967109.

39. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42(Database issue):D472–7. doi: 10.1093/nar/gkt1102 PMID: 24243840; PubMed Central PMCID: PMC3965010.

40. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel). 2012; 4(4):1180–211. doi: 10.3390/cancers4041180 PMID: 24213504; PubMed Central PMCID: PMC3712731.

41. Kollareddy M, Zheleva D, Dzubak P, Brahmkshatriya PS, Lepsik M, Hajduch M. Aurora kinase inhibitors: progress towards the clinic. Invest New Drugs. 2012; 30(6):2411–32. doi: 10.1007/s10637-012-9798-6 PMID: 22350019; PubMed Central PMCID: PMCPMC3484309.

42. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Research. 2010; 20(7):883–9. doi: 10.1101/gr.104695.109 WOS:000279404700002. PMID: 20418490

43. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell. 2014; 156(6):1324–35. doi: 10.1016/j.cell.2014.01.051 WOS:000332945100020. PMID: 24630730

44. Wu J, Guan X, Li Y-T, Bai P, Wu J. Matrix metalloproteinase7 -181A/G polymorphism is associated with increased cancer risk among high-quality studies: Evidence from a meta-analysis. Clinical Biochemistry. 2013; 46:1649–54. doi: 10.1016/j.clinbiochem.2013.07.015 PMID: 23895900

45. Didem T, Faruk T, Senem K, Derya D, Murat S, Murat G, et al. Clinical significance of serum tenascin-c levels in epithelial ovarian cancer. Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine. 2014; 35:6777–82. doi: 10.1007/s13277-014-1923-z

46. Rafii A, Halabi N, Malek J. High-prevalence and broad spectrum of Cell Adhesion and Extracellular Matrix gene pathway mutations in epithelial ovarian cancer. Journal of Clinical Bioinformatics. 2012; 2 (1). doi: 10.1186/2043-9113-2-15

47. Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X, et al. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell reports. 2013; 4(3):542–53. Epub 2013/08/13. doi: 10.1016/j.celrep.2013.07.010 PMID: 23933257.

48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352 PMID: 19505943; PubMed Central PMCID: PMC2723002.

50. Bauer D. Variant calling comparison CASAVA1.8 and GATK. Nature Precedings. 2011. doi: 10.1038/npre.2011.6107.1

51. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. Epub 2010/07/21. doi: 10.1101/gr.107524.110 PMID: 20644199; PubMed Central PMCID: PMC2928508.

52. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR}: ultrafast universal {RNA}-seq aligner. Bioinformatics. 2013; 29(1). doi: 10.1093/bioinformatics/bts635

53. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009; 25(17):2283–5. doi: 10.1093/bioinformatics/btp373 PMID: 19542151; PubMed Central PMCID: PMC2734323.

54. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr Protoc Bioinformatics. 2013; 44:15 4 1–4 7. doi: 10.1002/0471250953.bi1504s44 PMID: 25553206; PubMed Central PMCID: PMC4278659.

55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009; 25(16):2078–9. doi: 10.1093/bioinformatics/btp352

56. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015; 112(17):5473–8. doi: 10.1073/pnas.1418631112 PMID: 25827230; PubMed Central PMCID: PMC4418901.

57. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, 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. 2012; 6(2):80–92. Epub 2012/06/26. doi: 10.4161/fly.19695 PMID: 22728672.

58. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008; Chapter 10:Unit-10.1.

59. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–30. Epub 2013/11/15. doi: 10.1093/bioinformatics/btt656 PMID: 24227677.

60. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. doi: 10.1093/bioinformatics/btp616 PMID: 19910308

61. Schroeder M, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data. Genome medicine. 2013; 5(1). doi: 10.1186/gm413

62. Hulsen T, de Vlieg J, Alkema W. BioVenn—a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. Bmc Genomics. 2008; 9. Artn 488 doi: 10.1186/1471-2164-9-488 WOS:000261169600001. PMID: 18925949