MelanomaDB: a web tool for integrative analysis of melanoma genomic information to identify disease-associated molecular pathways

Alexander J. Trevarton[1], Michael B. Mann[2], Christoph Knapp[3], Hiromitsu Araki[4], Jonathan D. Wren[5,6], Steven Stones-Havas[7], Michael A. Black[8,9] and Cristin G. Print[1,3,9] *

[1] Department of Molecular Medicine and Pathology, School of Medical Sciences, University of Auckland, Auckland, New Zealand
[2] Cancer Research Program, The Methodist Hospital Research Institute, Houston, TX, USA
[3] Bioinformatics Institute, University of Auckland, Auckland, New Zealand
[4] Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, Fukuoka, Japan
[5] Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA
[6] Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA
[7] Biomatters Ltd., Auckland, New Zealand
[8] Department of Biochemistry, University of Otago, Dunedin, New Zealand
[9] Maurice Wilkins Centre, Auckland, New Zealand

Despite on-going research, metastatic melanoma survival rates remain low and treatment options are limited. Researchers can now access a rapidly growing amount of molecular and clinical information about melanoma. This information is becoming difficult to assemble and interpret due to its dispersed nature, yet as it grows it becomes increasingly valuable for understanding melanoma. Integration of this information into a comprehensive resource to aid rational experimental design and patient stratification is needed. As an initial step in this direction, we have assembled a web-accessible melanoma database, MelanomaDB, which incorporates clinical and molecular data from publically available sources, which will be regularly updated as new information becomes available. This database allows complex links to be drawn between many different aspects of melanoma biology: genetic changes (e.g., mutations) in individual melanomas revealed by DNA sequencing, associations between gene expression and patient survival, data concerning drug targets, biomarkers, druggability, and clinical trials, as well as our own statistical analysis of relationships between molecular pathways and clinical parameters that have been produced using these data sets. The database is freely available at http://genesetdb.auckland.ac.nz/melanomadb/about.html. A subset of the information in the database can also be accessed through a freely available web application in the Illumina genomic cloud computing platform BaseSpace at http://www.biomatters.com/apps/melanoma-profiler-for-research. The MelanomaDB database illustrates dysregulation of specific signaling pathways across 310 exome-sequenced melanomas and in individual tumors and identifies the distribution of somatic variants in melanoma. We suggest that MelanomaDB can provide a context in which to interpret the tumor molecular profiles of individual melanoma patients relative to biological information and available drug therapies.

Keywords: melanoma, mutation, molecular pathway, MelanomaDB, gene set analysis, BaseSpace

INTRODUCTION

THE GROWTH AND COMPLEXITY OF MELANOMA GENOMIC DATA

Melanoma researchers are faced with a rapidly growing amount of useful molecular and clinical data, particularly gene expression information. This rapid growth can be illustrated by surveying the Gene Expression Omnibus (GEO) (1), an international repository that contains a large subset of the published gene expression data (**Figure 1**). Largely based on genomic data, our understanding of the genes involved in melanoma progression has advanced from focused investigations of candidate genes to studies on a whole-genome scale (2). The advent of next-generation sequencing (NGS) in particular has opened up a floodgate of data, from the published sequence of the first melanoma genome in the beginning of 2010 (3), to more recent whole-exome studies sequencing more than one hundred tumors (4, 5). Melanoma genomic data is poised to grow rapidly with the advent of large-scale initiatives such as Australia's Melanoma Genome Project[1], melanoma analysis in The Cancer Genome Atlas (TCGA) project[2] as well as the melanoma sequencing projects underway at several individual institutions.

LIMITATIONS OF CURRENT TECHNIQUES

Unfortunately, information pertinent to melanoma exists in a diverse range of formats and locations. For example, relevant data
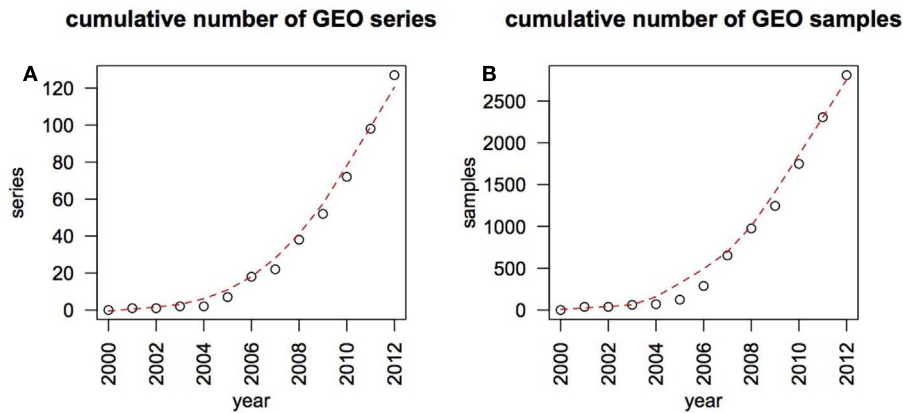
---

[1] http://www.melanoma.org.au/research/melanoma-genome-project.html
[2] http://cancergenome.nih.gov

FIGURE 1 | Growth of melanoma genomic data in the GEO database. The GEO database was searched on a year by year basis, using the MESH term "melanoma" and excluding records containing the phrase "cell line." By the end of January 2013 GEO contained 128 data series made up of 2819 samples that match these search criteria. The cumulative number of data series (submitted experiments) **(A)** and individual samples **(B)** are plotted as black circles, overlaid by a red trend line fitted over this data using the loess method.

about a single gene of interest may include information about the encoded protein's structure, cellular location, and function, contribution to molecular pathways, drugs that target the protein, the gene, or protein's utility as a biomarker, genome-wide association studies, mutation frequency, chromosomal aberrations, as well as RNA expression associations with metastasis, treatment response and patient survival, clinical SNP associations, and the results of literature mining. Even within the single data type of tumor DNA sequencing, a variety of methods have been used to implicate genes in melanoma initiation and progression, and these different methods produce data in differing formats. Ideally, all these diverse forms of data could be used by researchers in an integrated fashion to triangulate in on clinically important genes.

As a further challenge, genomic information in melanoma is particularly dense due to the high mutation rate found in melanomas of sun-exposed skin (6). This is likely to be due to both ultraviolet radiation-induced DNA damage and defects in DNA repair mechanisms (3). In addition, sequencing studies suggest that malignant melanoma is a relatively heterogeneous neoplasia with a range of driver mutations (5). Despite its potential value, coherent analysis of melanoma genomic information remains difficult for individual researchers. Data repositories such as Oncomine (7), Ingenuity Pathways Analysis[3], the Catalogue of Somatic Mutations in Cancer (COSMIC) (8, 9), and the Broad Institute's Melanoma Genomics Portal (10) bring together a massive amount of useful melanoma data. However, these disparate resources do not yet enable the full potential of integrated analysis of molecular pathways across different types of data associated with melanoma.

POTENTIAL CLINICAL USE OF MOLECULAR PATHWAY DATA ABOUT INDIVIDUAL TUMORS

Tumor development involves multiple genes encoding proteins and non-coding RNAs operating in molecular pathways.

Therefore, inference of molecular pathway activity from tumor genomic data using methods such as gene set analysis (GSA) (11) is useful in oncology (12, 13). Gene sets used for analysis may consist of co-expressed genes downstream of a specific molecular pathway (14) or genes that share common transcription factor binding sites (15). Statistical summaries of these gene sets have been used to infer molecular pathway activity, and these gene sets are frequently conserved across species (16). GSA has identified several molecular pathways associated with melanoma (17, 18), and can be used to identify the putative functional changes caused by the mutation, DNA gain or loss, and/or altered expression of genes in a particular patient's tumor. Popular GSA tools include GATHER (19), DAVID (20), GSEA (21), and GeneSetDB (22).

The number of clinically available targeted therapies for melanoma remains limited compared to the diverse genetic drivers of this tumor. Nevertheless, identification of drugs targeting a small number of melanoma drivers has been a major advance. For example, Vemurafenib targets the Mitogen Activated Protein Kinase (MAPK) pathway molecule BRAF (23). However, Vemurafenib is only indicated in *BRAF* V600E or V600K containing tumors and the majority of treated patients show relatively short term remission, with their relapse almost certainly caused by reactivation of the MAPK pathway, commonly through mutations in *NRAS* or *PDGFRB* (24). We propose that integration of molecular pathway data at both the patient population scale and individual tumor scale could help researchers better understand phenomena such as Vemurafenib resistance, and permit identification of rationally selected combinatorial therapies based on molecular stratification of patients.

EXPERIMENTAL OBJECTIVES

In the work described here, we have amalgamated a diverse range of genomic and clinical melanoma data, on the scales of both patient population and individual tumor into a single resource. This resource is provided as a downloadable file that can be searched and filtered using any spread sheet application. To facilitate use

---

[3]http://ingenuity.com/products/pathways_analysis.html

of this resource in the context of molecular pathways, we also provide a web-accessible SQL database named MelanomaDB, through which researchers can perform GSA using integrated melanoma data of several types. A subset of the information in the database can also be accessed through a freely available web application in the Illumina genomic cloud computing platform BaseSpace. While other disease-specific databases exist for other cancers such as lung (25) and ovarian (26) cancer, we know of no other database similar to ours dealing with melanoma. Furthermore, we believe that MelanomaDB's breadth across sequence and microarray data, biological and pharmacological gene sets, and pathway information, in addition to its usability and its melanoma focus, make it unique. In this paper, we use information assembled in MelanomaDB in several downstream analyses to demonstrate the utility of this resource for finding relationships between molecular pathways and clinical parameters, including the mutational patterns of members of molecular pathways (27) in individual tumors. We hope this tool will prove increasingly useful as it expands when new tumor data becomes available. In particular, we hope that it will provide a context in which to interpret the tumor molecular profiles of individual melanoma patients.

MATERIALS AND METHODS

OVERVIEW OF THE CONSTRUCTION OF MELANOMA GENE SETS

To facilitate an integrative analysis of melanoma information we combined a variety of melanoma data in the form of gene sets, attempting to collect information for all genes in the genome. These melanoma gene sets were groups of genes that shared biological or clinical relevance for melanoma, derived from five types of publically available information: drug and biomarker information, druggability, literature relationship strength, disease-specific survival, and somatic mutation data. Drug information includes information on compounds and the proteins they target, while Druggability information comprises of estimations of the degree to which proteins are amenable to targeting by drugs, and protein characteristics relevant to this. A detailed description of this information is available in Data Sheet 1 in Supplementary Material.

SOURCES OF SPECIFIC INFORMATION

Further explanations of the gene sets used are in the MelanomaDB help page at http://genesetdb.auckland.ac.nz/melanomadb/help.html

Drug and biomarker information

Drug information was taken from online databases DrugBank version 3 (28), KEGG DRUG (27), Therapeutic Targets Database (29), and ClinicalTrials.gov. Biomarker information was taken from published papers by Gould Rothberg et al. (30), Schramm and Mann (31), Utikal et al. (85), Mehta et al. (32), and from the database KEGG BRITE (27). It should be noted that gene sets such as those derived from DrugBank include all genes encoding proteins to which each drug binds, including both intended and unintended targets. However, metabolising enzymes, transporters and carrier proteins are excluded. For example, targets of the drug Cetuximab include the intended target (the human epidermal growth factor receptor) but also compliment components and Fc receptors, as is

expected due to the nature of this drug as an antibody[4]. For further explanations of the gene sets used see the MelanomaDB help page at http://genesetdb.auckland.ac.nz/melanomadb/help.html

Druggability information

Druggability data was sourced from the Sophic Integrated Druggable Genome Database (33), EBI's DrugEBIlity database (34), and published papers by Li and Lai (35) and Tiedemann et al. (36). Data on protein characteristics relevant to druggability were taken from Affymetrix annotations[5], and online databases UniProt Consortium (37), Secreted Protein Database (38), and KinBase (39).

Literature and genomic data relationship strength information

Information on Literature Relationship strength was derived from the IRIDESCENT (40) and GAMMA (41) software packages. IRIDESCENT searches every published MEDLINE abstract for associations between objects, and creates a network of tentative relationships between these objects. Objects encompass genes, diseases, phenotypes, chemical compounds, drugs, and ontology categories. The relative strength of association between two objects is determined by the frequency in which they appear in the same abstract or sentence. Here, this network is used to score the strength of association between genes and the terms "melanoma" or "metastatic melanoma."

GAMMA conducts a meta-analysis of gene expression behavior across 16,000 wide-ranging microarray experiments to identify genes that are consistently and specifically co-expressed across heterogeneous experimental conditions. In this way GAMMA extends the connections in IRIDESCENT's association network to genes without any published associations to melanoma by identifying which of these genes are consistently co-expressed with multiple known melanoma genes. To date, GAMMA has been used successfully to identify phenotypes and/or disease relevance for several previously uncharacterized genes (42–45).

Disease-specific survival data

Strength of statistical associations between RNA abundance and melanoma-specific survival were gathered from several published studies, and from our additional statistical analysis of two published sets of linked microarray and clinical data. Associations between gene expression in melanomas and patient survival were taken directly from John et al. (46), Mandruzzato et al. (47), and Journe et al. (48), and associations between gene expression and metastasis were taken directly from Timar et al. (49). We performed our own analyses on the microarray data of Bogunovic et al. (50) and Jönsson et al. (51) based on patient survival data and Affymetrix CEL files retrieved from GEO. The Bogunovic study's raw Affymetrix data was normalized using RMA normalization performed using the affy package in the R statistical software (52). The Illumina data from the Jonsson et al. study was obtained in a normalized format, however, we removed three patients for whom patient survival data was missing, and adjusted all microarray values by adding the minimum value in order to eliminate negative values. R was used to split the patients into two groups,

---

[4]http://www.drugbank.ca/drugs/DB00002
[5]http://www.affymetrix.com/support/technical/annotationfilesmain.affx

create a survival object for each group and then compare these two survival objects using a Log Rank test. For each probe set this splitting was performed nine times, once at each RNA abundance decile across the patient population. R was also used to fit a Cox proportional hazards regression model for each probe set.

To facilitate the use of these data in exploratory analyses for hypothesis generation, we also generated additional gene sets in which we aggregated several different RNA associations with patient survival to allow broader surveys. For example, four gene sets were identified from the expression and survival data of Bogunovic et al. (50) using different statistical criteria.

Somatic variant data

Multiple studies reporting melanoma variants were collated for use with MelanomaDB. A literature review identified 11 exome sequencing studies suitable for inclusion (4–6, 53–60). In addition, the Cancer Cell Line Encyclopedia (61), and the Sanger Institute's COSMIC (8, 9), and Matched Pair Cancer Cell Lines (3) were searched for mutations detected in melanoma cell lines. In total, we collected data on 58 established melanoma cell lines, 119 primary "short-passage" cell lines, 38 primary tumors, and 96 metastatic melanoma tumors. Non-silent variants were reported in 16,488 genes. With the exception of the 10 samples from the 2010 study of Berger et al. (53), and some of the samples from COSMIC, these samples have all been paired with matched normal samples to ensure that the variants reported are somatic. In the current iteration of this database only non-synonymous coding mutations, indels, splice-site mutations, and structural rearrangements (including gene fusions and read-through transcripts) are included. Synonymous coding mutations are not included. Presently, this somatic variant data includes more than 35,000 non-synonymous coding mutations, and more than 3,500 structural rearrangements and indels. We have not provided this somatic variant data as a supplementary file but instead invite readers to contact us to obtain the links to this data. We do this so we can ensure that access permission and ethical issues associated with this individual patient data are adhered to.

AMALGAMATION OF ALL DATA INTO GENE SETS

To facilitate the construction of gene sets, all data described above was combined into a single matrix, which is available as Data Sheet 2 in Supplementary Material. This matrix is gene-based and uses Entrez Gene ID as a unique index for each gene[6]. Every gene is represented by one row, and each column contains data from a single source. Columns annotating genes with references to other databases were derived from NCBI's Gene database FTP directory[7] and supplemented by Affymetrix annotations (see text footnote 5).

From this data matrix, a number of gene sets were derived. In most cases, columns of the matrix were converted directly into gene sets by including in that set every gene with an entry in that column. In some cases, such as statistical associations between RNA expression and patient survival, a cut-off was required for defining gene set membership. For example, only genes encoding proteins with positive DrugEBIlity ensemble scores were included

in the gene set "DrugEBIlity: Positive ensemble scores." A further description of the melanoma gene sets is available in Data Sheet 1 in Supplementary Material.

SQL DATABASE GENERATION

To facilitate access, combination, and filtering of different types of genomic data related to melanoma, and interpretation of this data in terms of molecular pathways and functional categories, the data matrix described above was used to generate a web-accessible SQL database named MelanomaDB. The web interface is implemented using Apache, PHP, Javascript, and HTML. The meta-gene set database GeneSetDB (22) was accessed from within MelanomaDB to identify the intersection between melanoma-specific gene sets and gene sets related to biological functions and molecular pathways. The R framework was used for statistical calculations. GSA was performed using the hyper-geometric distribution to calculate the probability of overrepresentation, followed by multiple testing correction using the Benjamini and Hochberg method (62).

BaseSpace APPLICATION PREPARATION

A subset of the information in MelanomaDB is also included in a freely available Illumina BaseSpace application. This BaseSpace application retrieves a tumor and corresponding normal germ line sequence pair from the BaseSpace archive or the user's own BaseSpace account as vcf files. Then, variants present in the tumor but the not normal germ line tissue of the patient are identified using the Genome Analysis Tool Kit's SelectVariants java tool (63). This list of tumor variant genes is identified. Then, the molecular pathways these genes correspond to, along with any statistically significant pathway enrichment within the list of variant genes and targeting drugs, are retrieved from the GeneSetDB pathway analysis web tool (22). A diagram showing tumor variant genes in the context of molecular pathways is generated using the KEGG, Reactome, and Biocarta pathways included in the R graphite package (64), and a clustered heatmap showing how the genetic variants in the sample tumor compare to variants in the 310 tumors cataloged in MelanomaDB is generated. This clustered heatmap is generated: (i) using a modification of the heatmap.2 function from the R gtools package (see Data Sheet 5 in Supplementary Material) (65), using the "binary" method for distance calculation and the "single" method for clustering and (ii) as a reverse-orientation waterfall plot to illustrate patterns of somatic variant co-occurrence in melanoma.

ASSEMBLY OF INFORMATION FOR INDIVIDUAL TUMORS

From the exome and whole-genome sequencing information assembled above, we constructed a tumor-based matrix in which each row was a gene, each column was an individual tumor and each cell described any somatic variants present in a certain gene for a certain tumor. After duplicated tumors were removed, this somatic variant data included 310 samples, 183, and 72 of which had somatic alterations in the BRAF and NRAS genes, respectively. When multiple sequenced tumors or cell lines from the same patient were available, the union of somatic variants found in these samples was used. Links to the papers and their supplementary web sites used to construct this tumor-specific somatic variant data is available in Data Sheet 3 in Supplementary Material. The authors

---

[6]http://www.ncbi.nlm.nih.gov/gene, accessed on 30/7/12
[7]ftp://ftp.ncbi.nih.gov/gene/DATA/

can assist researchers with the precise sources of information used to construct this resource.

VISUALIZATION

The statistical software R was used to construct a clustered heat map of tumor variants for genes included in the KEGG "Melanoma" signalling pathway with a modified heatmap.2 function of the R package "gplots,"[8] using the "binary" method for distance and the "single" method for clustering. R was also used to draw gene network diagrams. Molecular pathways were obtained from the pathways included in the graphite R package[9] and were plotted using the graphite (see text footnote 9) R package.

The R scripts used to generate **Figures 2A–C** as well as the pathway diagrams and heatmaps in **Figures 4–7** are given in Data Sheet 5 in Supplementary Material.

RESULTS AND DISCUSSION

Here we describe the assembly and use of the MelanomaDB database.

ASSEMBLY OF MELANOMA GENOMIC INFORMATION FROM DIVERSE SOURCES INTO A MELANOMA DATA MATRIX

Firstly, a melanoma data matrix (Data Sheet 2 in Supplementary Material) was constructed, with genes (or genomic loci in some cases) as rows. The columns of this matrix represent diverse features of biological functions related to melanoma and are described in Data Sheet 1 in Supplementary Material. This melanoma data matrix can be utilized in a variety of ways. Most simply, researchers can access a variety of data pertaining to their particular gene of interest. The melanoma data matrix can also be manipulated with spread sheet software to sort, find, and filter information in order to generate gene lists useful for hypothesis generation.

ASSEMBLY OF SOMATIC VARIANT INFORMATION FOR MELANOMAS OF INDIVIDUAL PATIENTS

Next, we assembled as much information about somatic variation in individual exome-sequenced and genome-wide-sequenced melanomas as possible. We gathered information about somatic variations in 58 established melanoma cell lines, 119 primary "fresh" cell lines, 38 primary tumors, and 96 metastatic melanoma tumors, which was appended to the information matrix described above (Data Sheet 3 in Supplementary Material, Tab "Tables Used"). Information about non-synonymous coding mutations, structural rearrangements, and indels was included (intronic and synonymous coding mutations were excluded from the current iteration of this data resource). The information contained in Data Sheet 2 in Supplementary Material was read into the statistical environment R and visualized, as described in the Section "Materials and Methods" and Data Sheet 5 in Supplementary Material. Firstly, the distribution of somatic variations for individual genes is shown in **Figure 2A**. The majority of genes showed somatic variations in only small numbers of tumors. Comparison of each gene's total exon length versus the number of tumors with a mutation in
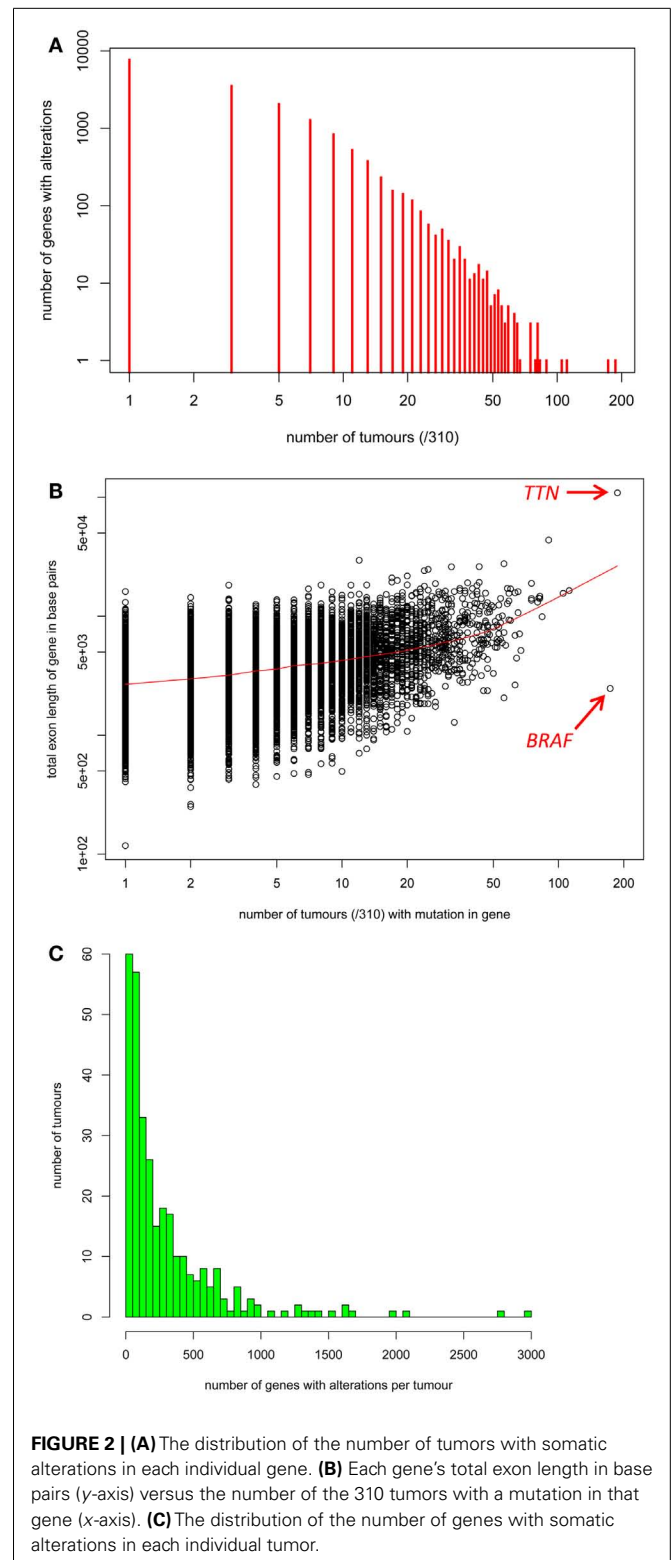
FIGURE 2 | (A) The distribution of the number of tumors with somatic alterations in each individual gene. **(B)** Each gene's total exon length in base pairs (y-axis) versus the number of the 310 tumors with a mutation in that gene (x-axis). **(C)** The distribution of the number of genes with somatic alterations in each individual tumor.

that gene using R (**Figure 2B**), revealed a statistically significant but weak correlation between somatic variation frequency and total exon length (Pearson's correlation coefficient $= 0.47$, $p \leq 0.001$). Although variations in large genes such as *Titan* (*TTN*) have been

implicated as cancer drivers, these may also occur in so many melanomas due to large gene size increasing the likelihood of passenger mutations. However, the *BRAF* gene clearly stands out as frequently mutated in melanomas despite its moderate length. The distribution of the number of genes with somatic alterations in each individual tumor was performed using R and is shown in **Figure 2C**.

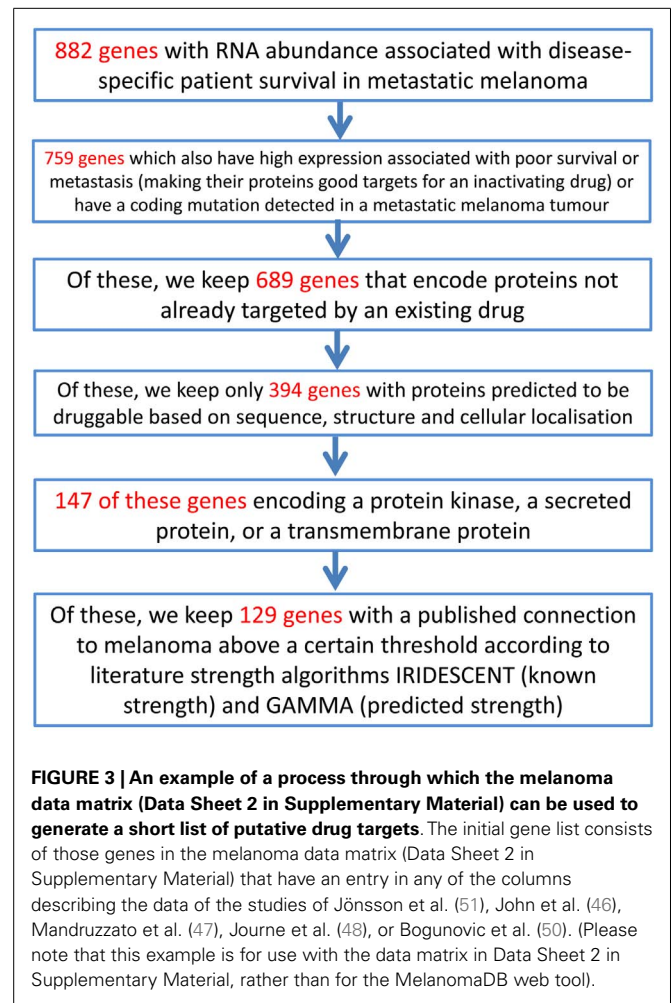### USE OF THE COMBINED MELANOMA INFORMATION

As an example of using the information assembled above, an approach to identifying novel candidate novel drug targets for melanoma using this melanoma data matrix (Data Sheet 2 in Supplementary Material) can be performed by filtering and sorting Data Sheet 2 in Supplementary Material in a spreadsheet application and is described in **Figure 3**.

This process generates a short list of 129 genes that can be examined more closely in order to select a final list of genes that may warrant investigation in the laboratory. A variant on this approach may be to place more weight on particular data, for example, on selected druggability measures. By using a spreadsheet application to take the 987 genes in Data Sheet 2 in Supplementary Material encoding proteins that have scored greater than 0.5 on either DrugEBIlity's Ensemble score or Li and Lai's druggability measure, and eliminating proteins already targeted by existing drugs, we have a list of 803 genes that are predicted to be probably druggable. Of these, 21 also have high RNA expression significantly associated with reduced disease-free survival in melanoma patients, making them possible new drug targets. These genes are *AKR7A2, AKR7A3, ARIH1, ARPC1A, CD163, DCT, DHRS11, DUS4L, FAH, FSCN1, HS3ST3A1, NRAS, NUP155, PANK2, PRMT3, QTRT1, RAD1, RAE1, SUV39H2, UPP1, USP13*. It is interesting to see *NRAS* on this list, which is a potential melanoma drug candidate but has proved remarkably resistant to drug development efforts to date (66). *CD163* expression on melanoma-infiltrating macrophages has been suggested as a prognostic marker in melanoma (67).

Similarly, a list of putative melanoma tumor suppressor genes or melanoma oncogenes can be generated using a spreadsheet application from this melanoma data matrix (Data Sheet 2 in Supplementary Material). For example, a list consisting of genes that are mutated in more than 10% of melanoma metastases and have shorter melanoma-free patient survival associated with their low (putative tumor suppressor) or high (oncogene) RNA expression. Known tumor suppressors and oncogenes that were identified by this strategy (*NRAS, KIT*, and *WNT* family members) were removed. This list of putative melanoma tumor suppressors and oncogenes that remains is shown in **Table 1**.

Combined melanoma information with gene set analysis

Combining this assembled melanoma information with statistical GSA can potentially provide additional insights. For example, with a spreadsheet application we could generate a list of 245 genes from Data Sheet 2 in Supplementary Material that have coding region mutations in more than 10% of melanoma metastases, and subject this list to gene set enrichment analysis in order to identify biological functions that may be commonly disrupted in melanoma. When submitted to the web tool GeneSetDB (a meta-database of biologically relevant sets of genes) for enrichment analysis (with

FIGURE 3 | An example of a process through which the melanoma data matrix (Data Sheet 2 in Supplementary Material) can be used to generate a short list of putative drug targets. The initial gene list consists of those genes in the melanoma data matrix (Data Sheet 2 in Supplementary Material) that have an entry in any of the columns describing the data of the studies of Jönsson et al. (51), John et al. (46), Mandruzzato et al. (47), Journe et al. (48), or Bogunovic et al. (50). (Please note that this example is for use with the data matrix in Data Sheet 2 in Supplementary Material, rather than for the MelanomaDB web tool).

false discovery rate set to 0.01), this list of 245 genes was found to be significantly enriched for several gene sets including sets associated with the extracellular matrix (ECM), cell adhesion, and collagen fibril organization. We encourage users to use a spreadsheet application and simple web tools such as GeneSetDB to perform their own exploration of Data Sheet 2 in Supplementary Material.

ASSEMBLY OF MelanomaDB – A WEB-ACCESSIBLE GENOMIC MELANOMA SQL DATABASE, AND OF A CORRESPONDING BaseSpace APP

In order to make use of this assembly of melanoma information and its regular updating easier, we converted this melanoma data matrix (Data Sheet 2 in Supplementary Material) into a web-accessible SQL database. This database, named MelanomaDB, features melanoma gene sets derived from Data Sheet 2 in Supplementary Material and directly links into a molecular pathway/GSA meta-database previously generated by our research group named GeneSetDB (22). Using MelanomaDB, a user can easily find the union or intersection between any number of melanoma gene sets (taken from the columns of Data Sheet 2 in Supplementary Material) and also their own user-submitted gene lists (copied and pasted, or uploaded from a file, using any of over 50 types of commonly used gene identifier), then interrogate the molecular

Table 1 | Four putative melanoma oncogenes and two putative tumor suppressor genes derived from the amalgamated data.

Entrez gene	Gene symbol	Gene title	Chromosomal location	Putative tumor suppressor or oncogene?
7373	COL14A1	Collagen, type XIV, alpha 1	8q23	Tumor suppressor
387357	THEMIS	Thymocyte selection associated	6q22.33	Tumor suppressor
6299	SALL1	Sal-like 1 (Drosophila)	16q12.1	Oncogene
5069	PAPPA	Pregnancy-associated plasma protein A, pappalysin 1	9q33.2	Oncogene
26278	SACS	Spastic ataxia of Charlevoix-Saguenay (sacsin)	13q12	Oncogene
81832	NETO1	Neuropilin (NRP) and tolloid (TLL)-like 1	18q22.2	Oncogene

pathways for which the genes in these lists are enriched. Multiple iterations are possible, so that a user might find the union of some melanoma-associated gene sets and then find the intersection of this union with other gene sets, which can finally be directly piped into the gene set meta-database GeneSetDB to identify enriched molecular pathways. MelanomaDB is available at http://genesetdb.auckland.ac.nz/melanomadb/about.html

A subset of the information in MelanomaDB was also included in a freely available Illumina BaseSpace application, which can be accessed at http://www.biomatters.com/apps/melanoma-profiler-for-research (click on "sample project" and navigate using green tabs at top of screen). This BaseSpace application performs variant calling against reference sequences for a user-defined tumor, then uses information from MelanomaDB to identify molecular pathways that genes which contain non-synonymous variants constitute. These pathways are visualized relative to targeting drugs and other clinically related information using pathway diagrams, heatmaps, and waterfall plots, in comparison to the 310 melanomas described above. We hope that this app may be of particular use to researchers involved in generating new melanoma tumor sequences.

MelanomaDB FACILITATES ASSESSMENT OF FUNCTIONAL RELATIONSHIPS INHERENT IN TUMOR SOMATIC VARIANTS

The tumor gene sequence information included in MelanomaDB allows calculation of the proportion of melanomas that carry somatic variations in each gene/loci on a genome-wide scale. For example, by selecting gene sets using the MelanomaDB web tool, we identified those genes in which over 10% of the 96 sequenced metastatic melanomas currently in the database carried non-synonymous somatic variations. This list of 245 genes included genes that have been the focus of recent publications describing mutations in melanoma, such as *PREX2* (6), *GRM3* (57), and *ERBB4* (56) [other melanoma-associated genes such as *MAP3K5/9* (58), *MAP2K1/2* (54), and *RAC1* (4–6) are included as mutated genes in human tumors in MelanomaDB but fall outside this list of 245 genes]. As would be expected, this composite list featured genes also indicated as frequently mutated in melanoma by the larger sequencing studies (4, 5) that were used in its construction, for example, half of the genes identified by Berger et al. (6) as "significantly mutated" appear on our composite list. By selecting the option in MelanomaDB to pipe these 245 genes to the GeneSetDB web tool, we identified that these genes were significantly enriched for a small group of biological functions including cell adhesion, collagen fibril organization, and ECM. Cell adhesion is
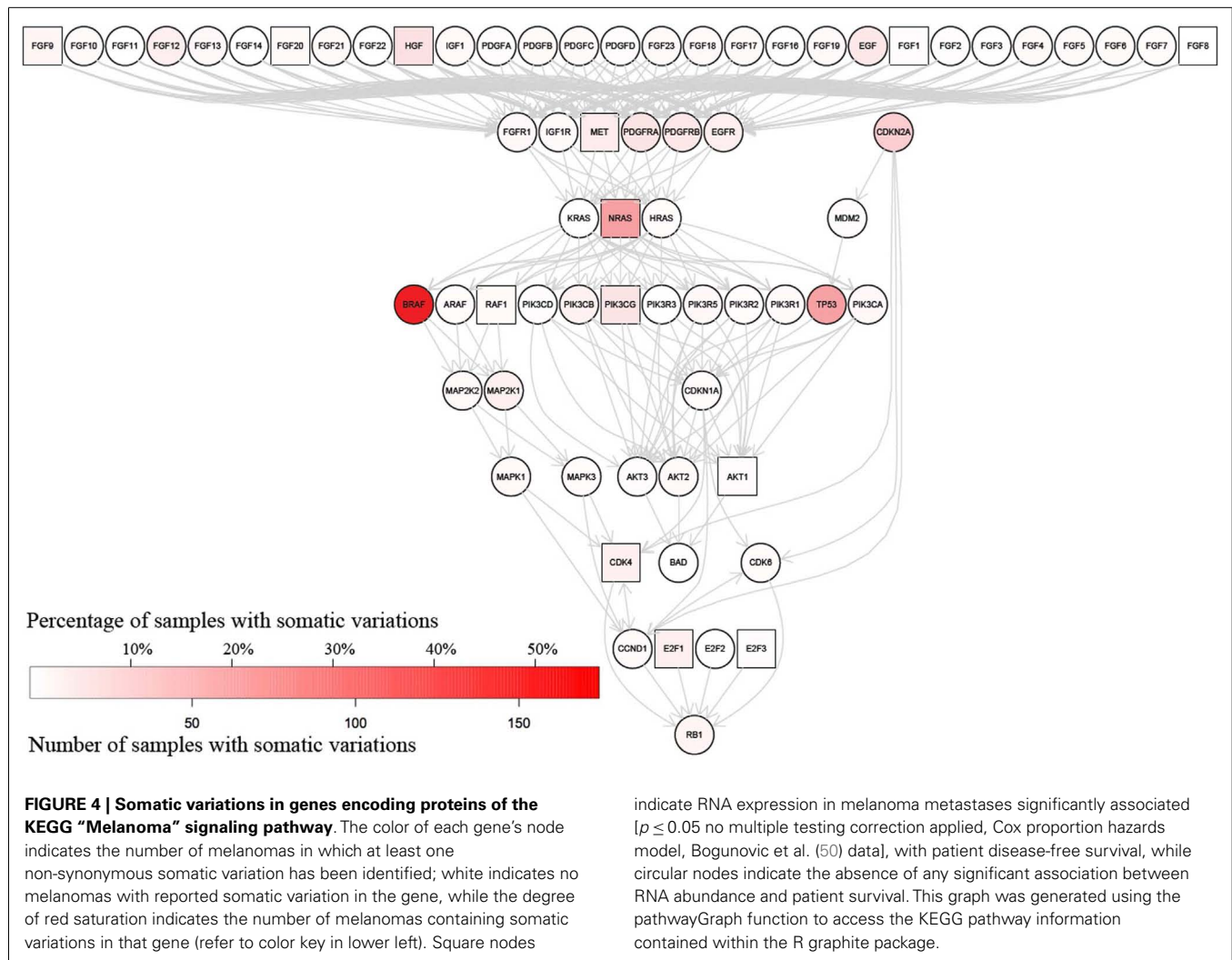
briefly mentioned in some of the sequencing studies' discussions (4, 54), and the ECM is a focus for one study (55). However, other pathways emphasized by these sequencing studies, such as the glutamate pathway (60) or chromatin remodeling pathways (5), did not feature in the results of our analysis.

ANALYSIS OF SPECIFIC SIGNALING PATHWAYS RELEVANT TO MELANOMA

The information in MelanomaDB can be used to annotate the signalling pathways contained within the R graphite package (27). This can be done either as a function of the MelanomaDB web tool, or using R scripts supplied in Data Sheet 5 in Supplementary Material. For example, **Figure 4** shows the KEGG pathway named "Melanoma" with nodes colored in shades of red according to the frequency of non-synonymous somatic variations. Thirteen nodes were plotted as boxes rather than circles to indicate that the abundance of their encoded mRNA in melanoma metastases was significantly associated with patient survival in our analysis of the data of Bogunovic et al. (50) (Cox proportional hazards model, $p \leq 0.05$, no multiple testing correction applied). Significantly more of the genes in the KEGG pathway named "Melanoma" carried more somatic variants than expected due to chance alone (Fisher's exact test with right-tailed hyper-geometric distribution, $p \leq 0.002$), in agreement with the known importance of the signaling events represented in this pathway to melanoma formation and progression.

ANALYSIS OF MELANOMA SIGNALING PATHWAYS IN INDIVIDUAL TUMORS

As an example of how this pathway-specific information can be used to place the tumors of individual patients into the context of tumors from the patient population, as well as into the context of other information within MelanomaDB, we used the information assembled here to draw a clustered heat map for genes encoding molecules of the KEGG "Melanoma" signaling pathway (**Figure 5**). This clustered heatmap is annotated with gene-survival associations, druggability indices, current drug targets, COSMIC census genes, known melanoma driver mutations and somatic variant frequency in melanoma. This can be done either as a function of the MelanomaDB web tool, or using R scripts supplied in Data Sheet 5 in Supplementary Material. In this analysis, somatic variants in genes drive the tumor clustering and potentially stratify patients into those with common biological changes, which may be susceptible to particular pathway-targeted therapies. For instance, there is a cluster of tumors with *BRAF* as the

FIGURE 4 | Somatic variations in genes encoding proteins of the KEGG "Melanoma" signaling pathway. The color of each gene's node indicates the number of melanomas in which at least one non-synonymous somatic variation has been identified; white indicates no melanomas with reported somatic variation in the gene, while the degree of red saturation indicates the number of melanomas containing somatic variations in that gene (refer to color key in lower left). Square nodes indicate RNA expression in melanoma metastases significantly associated [$p \leq 0.05$ no multiple testing correction applied, Cox proportion hazards model, Bogunovic et al. (50) data], with patient disease-free survival, while circular nodes indicate the absence of any significant association between RNA abundance and patient survival. This graph was generated using the pathwayGraph function to access the KEGG pathway information contained within the R graphite package.
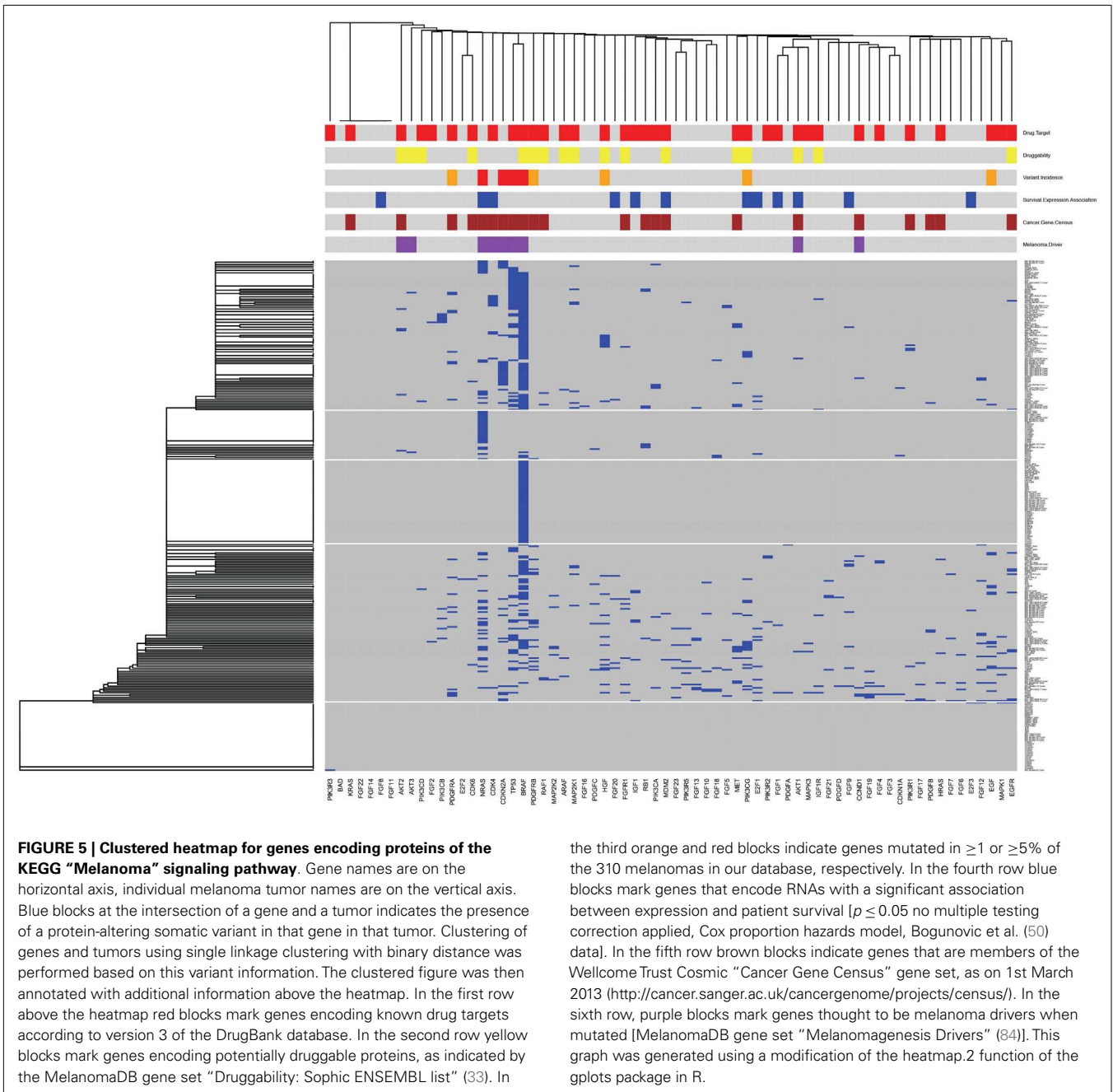
only somatic variant in this pathway (middle horizontal block in **Figure 5**). Of these 51 *BRAF*-variant only melanomas, 42 carry the *BRAF* V600E mutation and may putatively be tumors for stratification to Vemurafenib therapy, given their lack of somatic variants in genes encoding other proteins in this signaling pathway that could potentially contribute to Vemurafenib resistance. Some tumors carry only *NRAS* mutations, while others have either more complex mutational patterns, or no somatic mutations in this pathway. This is in accordance with previous studies reporting that mutations in *NRAS* and *BRAF* tend to be mutually exclusive but collectively occur in approximately 90% of melanomas (68). To assist interpretation of the different mutations seen in each tumor and in clusters of genetically similar tumors, the heatmap has been annotated with information about inferred melanoma driver mutations, known drug targets, and potentially druggable proteins. This type of heat map can be generated for any molecular pathway or combination of pathways. Extending this analysis, a new patient's mutation profile could be added to an established clustering analysis of large numbers of melanomas in order to identify which previously studied tumors were similar in mutation complement, which may assist prognostication and treatment stratification. In the future it will be interesting to use MelanomaDB to investigate the genomes of multiple samples from single melanomas to assess the intra-tumoral heterogeneity seen in this disease (69).

In addition, using a function in the MelanomaDB web tool of the R scripts supplied in Data Sheet 5 in Supplementary Material, somatic alteration of genes in specific molecular pathways can be drawn on a patient-by-patient basis (**Figure 6**). This allows visualization of protein-altering gene sequence variants in the context of the encoded protein's position in molecular pathways relevant to specific targeted therapies. For instance, using a well-known example from other tumor types, the position in pathway diagrams of a genetic variant known to be activating (e.g., mutant *KRAS*), downstream of a drug (e.g., cetuximab) target (e.g., EGFR) may indicate potential for resistance to the drug.

We then used an R script (Data Sheet 5 in Supplementary Material) to perform gene set enrichment analysis using the GATHER
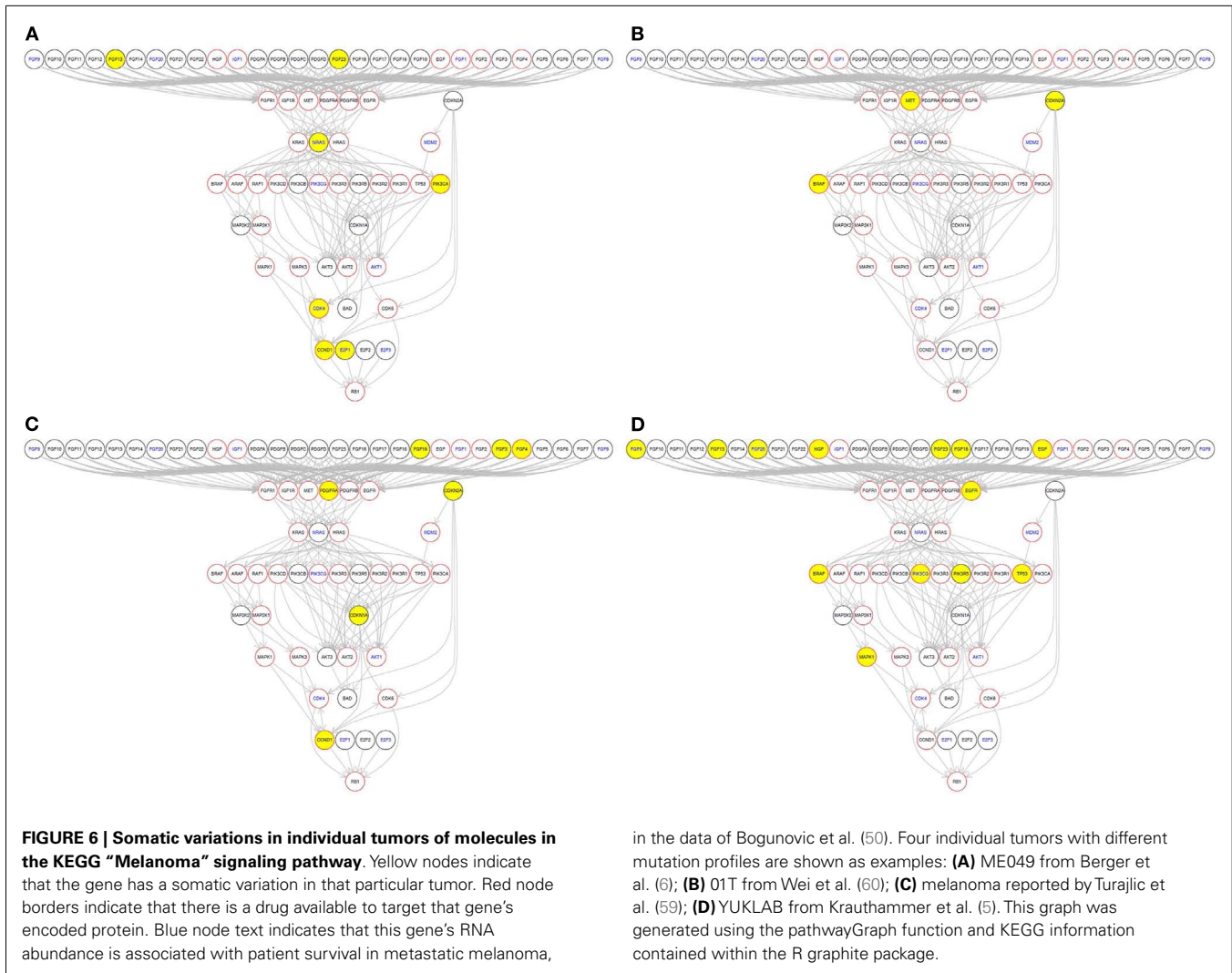
FIGURE 5 | Clustered heatmap for genes encoding proteins of the KEGG "Melanoma" signaling pathway. Gene names are on the horizontal axis, individual melanoma tumor names are on the vertical axis. Blue blocks at the intersection of a gene and a tumor indicates the presence of a protein-altering somatic variant in that gene in that tumor. Clustering of genes and tumors using single linkage clustering with binary distance was performed based on this variant information. The clustered figure was then annotated with additional information above the heatmap. In the first row above the heatmap red blocks mark genes encoding known drug targets according to version 3 of the DrugBank database. In the second row yellow blocks mark genes encoding potentially druggable proteins, as indicated by the MelanomaDB gene set "Druggability: Sophic ENSEMBL list" (33). In

the third orange and red blocks indicate genes mutated in ≥ 1 or $\geq 5\%$ of the 310 melanomas in our database, respectively. In the fourth row blue blocks mark genes that encode RNAs with a significant association between expression and patient survival [$p \leq 0.05$ no multiple testing correction applied, Cox proportion hazards model, Bogunovic et al. (50) data]. In the fifth row brown blocks indicate genes that are members of the Wellcome Trust Cosmic "Cancer Gene Census" gene set, as on 1st March 2013 (http://cancer.sanger.ac.uk/cancergenome/projects/census/). In the sixth row, purple blocks mark genes thought to be melanoma drivers when mutated [MelanomaDB gene set "Melanomagenesis Drivers" (84)]. This graph was generated using a modification of the heatmap.2 function of the gplots package in R.

web tool[10] (19) to identify any KEGG pathways for which genes somatically altered in each tumor were significantly enriched (Data Sheet 4 in Supplementary Material). KEGG pathways that appeared as significantly enriched in individual tumors included the "ECM receptor interaction" and "Neuroactive ligand-receptor interaction" KEGG pathways. To illustrate this, we selected one sequenced metastatic melanoma, ME029 from the Berger et al. (6) cohort, and drew these two pathways along with the KEGG "Melamoma" pathway for this single tumor (**Figure 7**). Two of

these pathways are drawn for all 310 tumors included in this study in: Presentation 1 ("Melanoma") and Presentation 2 ("Neuroactive ligand-receptor interaction").

LIMITATIONS OF OUR APPROACH

The approach we have described, while already functioning in a useful way as a melanoma-focused integrated genomic database, provides a template for further development to address the limitations below: (i) It will be important to identify the likely effects of specific somatic variations in the sequenced tumors (e.g., loss of function, altered function, or activation of the encoded protein). In future iterations of MelanomaDB, based on larger

[10]http://gather.genome.duke.edu

FIGURE 6 | Somatic variations in individual tumors of molecules in the KEGG "Melanoma" signaling pathway. Yellow nodes indicate that the gene has a somatic variation in that particular tumor. Red node borders indicate that there is a drug available to target that gene's encoded protein. Blue node text indicates that this gene's RNA abundance is associated with patient survival in metastatic melanoma, in the data of Bogunovic et al. (50). Four individual tumors with different mutation profiles are shown as examples: **(A)** ME049 from Berger et al. (6); **(B)** 01T from Wei et al. (60); **(C)** melanoma reported by Turajlic et al. (59); **(D)** YUKLAB from Krauthammer et al. (5). This graph was generated using the pathwayGraph function and KEGG information contained within the R graphite package.

numbers of tumors, we will include capacity to dissect the type of genetic alteration such as deletions, coding region mutations, promoter mutations, etc. The database may also be expanded to include the results of analyses from software that predict the effects of coding variants on protein function, such as SIFT (70), PolyPhen (71), or PROVEAN (72), as well as the known effects of specific mutations using resources such as COSMIC (8). (ii) Data on naevi and synonymous mutations can also be added. (iii) Information from model organisms such as mouse could also be added. (iv) Results from the ENCODE project (73) could be added along with whole genome sequencing of melanomas will allow inclusion of numerous additional functional genetic loci [e.g., ncRNAs, both general (74) and melanoma specific (75)] in the database. The ENCODE project suggests that mutations in regulatory regions such as distal enhancers can affect the expression of genes located hundreds of kilobases away (76); a way to include this in MelanomaDB could be to take a gene network approach to identify distant genes that have expression correlated with these mutations, as well as methods such as chromatin conformation capture (77). (v) Future additions to the database will

also aim to incorporate data concerning the role of epigenetics, including methylation, in melanoma (78–80). (vi) There is also room to expand upon melanoma drivers, such as those highlighted in GISTIC (81), JISTIC (82), and CONEXIC (83). (vii) There is an inherent risk in any assembly or meta-analysis of data from several sources that errors in the original data are perpetuated. While it is possible that the intersection of multiple independent sources of similar types of information may reduce the change of propagating random errors, systematic errors co-occur in independent data sources. This risk affects any project of this sort and is difficult to control. Here we have attempted to minimize this risk by selecting constituent databases that are extensively used and have been peer reviewed, and on which we could perform spot checks. We consider these data sources to be the best possible choices, within our ability to assess them. (viii) The final limitation is that the molecular pathways used when assembling this database are limited by current knowledge, and overlap with one another. The database will be updated with new pathway information as it becomes available. Identifying the pathways that are not affected can be as useful as identifying those that are. The data we have
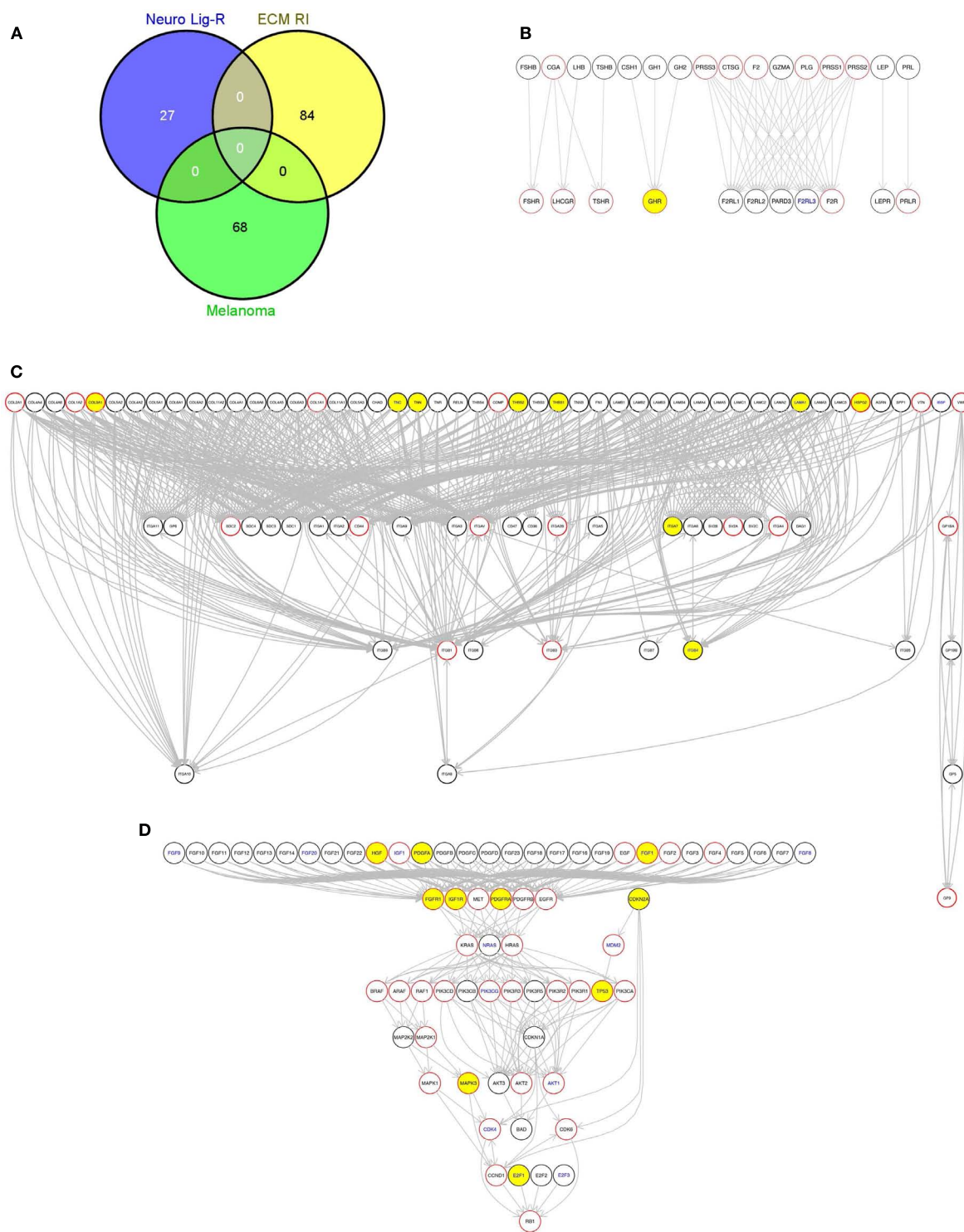
FIGURE 7 | (A) Venn diagram showing the overlap of genes between these three pathways used in this figure, generated using the Venny web tool (http://bioinfogp.cnb.csic.es/tools/venny/index.html). "Melanoma," "Neuro Lig-RI," and "ECM RI" indicate members of the "Melanoma," "Neuroactive ligand-receptor interaction," and "Extracellular matrix (ECM) receptor interaction" KEGG pathways, respectively, contained in the R graphite package; (B) The KEGG "Neuroactive ligand-receptor interaction" pathway; (C) The KEGG "Extracellular matrix (ECM) receptor interaction pathway"; (D) The KEGG "Melanoma" pathway. Yellow fill color in nodes indicate genes with protein-altering somatic variations in this sample. Nodes with red borders represent genes that encode targets of existing drugs according to version 3 of the DrugBank database. Nodes with blue text indicate genes that encode RNAs with a significant association between expression and patient survival [$p \leq 0.05$ no multiple testing correction applied, Cox proportion hazards model, Bogunovic et al. (50) data, see Materials and Methods]. This graph was generated using the pathwayGraph function and KEGG information contained within the R graphite package. Similar graphs can also be generated using the MelanomaDB web tool.

generated using literature relationships with the IRREDESCENT and GAMMA methods has not been experimentally verified and is intended primarily for hypothesis generating.

CONCLUSION

We have brought together a large collection of melanoma genomic data of several types from published studies and publicly available datasets into an easily utilized data matrix that can be analyzed using a spread sheet application. We also assembled data on tumors from individual patients. We then incorporated this information into a web-accessible SQL database, MelanomaDB, which researchers can use to perform molecular pathway and GSA of melanoma genomic data, and into a BaseSpace application. By way of illustration, we used this information to analyze the mutational and expression patterns of genes encoding proteins in specific directional signaling pathways within individual tumors, and annotated these visualizations with information about existing drugs, druggability, associations between RNA expression and survival, and driver mutations. We hope that this resource will prove increasingly useful when it expands as new tumor data becomes available. In particular, we hope it may provide a context in which to interpret the melanoma molecular profiles of new patients as

well as patient-specific molecular pathway disruption. We have demonstrated possible uses of this integrated information, and encourage melanoma researchers to employ these resources.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Cancer_Genetics/10.3389/fonc.2013.00184/abstract

REFERENCES

1. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* (2009) **37**(Database issue):D885–90. doi:10.1093/nar/gkn764

2. Walia V, Mu EW, Lin JC, Samuels Y. Delving into somatic variation in sporadic melanoma. *Pigment Cell Melanoma Res* (2012) **25**(2):155–70. doi:10.1111/j.1755-148X.2012.00976.x

3. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* (2010) **463**(7278):191–6. doi:10.1038/nature08658

4. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. *Cell* (2012) **150**(2):251–63. doi:10.1016/j.cell.2012.06.024

5. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, McCusker JP, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* (2012) **44**(9):1006–14. doi:10.1038/ng.2359

6. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* (2012) **485**(7399):502–6. doi:10.1038/nature11071

7. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* (2007) **9**(2):166–80. doi:10.1593/neo.07112

8. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* (2011) **39**(Database issue):D945–50. doi:10.1093/nar/gkq929

9. 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) **57**:10.11.1–10.11.26. doi:10.1002/0471142905.hg1011s57

10. Lin WM, Baker AC, Beroukhim R, Winckler W, Feng W, Marmion JM, et al. Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Res* (2008) **68**(3):664–73. doi:10.1158/0008-5472.CAN-07-2615

11. Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics* (2008) **9**:502. doi:10.1186/1471-2105-9-502

12. Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci U S A* (2012) **109**(8):2724–9. doi:10.1073/pnas.1018854108

13. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, et al. A pathway-based classification of human breast cancer. *Proc Natl Acad Sci U S A* (2010) **107**(15):6994–9. doi:10.1073/pnas.0912708107

14. Obayashi T, Kinoshita K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res* (2009) **16**(5):249–60. doi:10.1093/dnares/dsp016

15. Hatanaka Y, Nagasaki M, Yamaguchi R, Obayashi T, Numata K, Fujita A, et al. A novel strategy to search conserved transcription factor binding sites among coexpressing genes in human. *Genome Inform* (2008) **20**:212–21. doi:10.1142/9781848163003_0018

16. Obayashi T, Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res* (2011) **39**(Database issue):D1016–22. doi:10.1093/nar/gkq1147

17. Hoek K, Rimm DL, Williams KR, Zhao H, Ariyan S, Lin A, et al. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Res* (2004) **64**(15):5270–82. doi:10.1158/0008-5472.CAN-04-0731

18. Wang L, Hurley DG, Watkins W, Araki H, Tamada Y, Muthukaruppan A, et al. Cell cycle gene networks are associated with melanoma prognosis. *PLoS ONE* (2012) **7**(4):e34247. doi:10.1371/journal.pone.0034247

19. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* (2006) **22**(23):2926–33. doi:10.1093/bioinformatics/btl483

20. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* (2009) **4**(1):44–57. doi:10.1038/nprot.2008.211

21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* (2005) **102**(43):15545–50. doi:10.1073/pnas.0506580102

22. Araki H, Knapp C, Tsai P, Print C. GeneSetDB: a comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio* (2012) **2**:76–82. doi:10.1016/j.fob.2012.04.003

23. Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P, Nolop K, et al. Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat Rev Drug Discov* (2012) **11**(11):873–86. doi:10.1038/nrd3847

24. Nazarian R, Shi H, Wang Q, Kong X, Koya RC, Lee H, et al. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* (2010) **468**(7326):973–7. doi:10.1038/nature09626

25. Wang L, Xiong Y, Sun Y, Fang Z, Li L, Ji H, et al. HLungDB: an integrated database of human lung cancer research. *Nucleic Acids Res* (2010) **38**(Database issue):D665–9. doi:10.1093/nar/gkp945

26. Kaur M, Radovanovic A, Essack M, Schaefer U, Maqungo M, Kibler T, et al. Database for exploration of functional context of genes implicated in ovarian cancer. *Nucleic Acids Res* (2009) **37**(Database issue):D820–3. doi:10.1093/nar/gkn593

27. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* (2012) **40**(Database issue):D109–14. doi:10.1093/nar/gkr988

28. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res* (2011) **39**(Database issue):D1035–41. doi:10.1093/nar/gkq1126

29. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* (2012) **40**(Database issue):D1128–36. doi:10.1093/nar/gkr797

30. Gould Rothberg BE, Bracken MB, Rimm DL. Tissue biomarkers for prognosis in cutaneous melanoma: a systematic review and meta-analysis. *J Natl Cancer Inst* (2009) **101**(7):452–74. doi:10.1093/jnci/djp038

31. Schramm SJ, Mann GJ. Melanoma prognosis: a REMARK-based systematic review and bioinformatic analysis of immunohistochemical and gene microarray studies. *Mol Cancer Ther* (2011)

32. Mehta S, Shelling A, Muthukaruppan A, Lasham A, Blenkiron C, Laking G, et al. Predictive and prognostic molecular markers for cancer medicine. *Ther Adv Med Oncol* (2010) **2**(2):125–48. doi:10.1177/1758834009360519

33. Blake PM, Decker DA, Glennon TM, Liang YM, Losko S, Navin N, et al. Toward an integrated knowledge environment to support modern oncology. *Cancer J* (2011) **17**(4):257–63. doi:10.1097/PPO.0b013e31822c390b

34. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* (2012) **40**(Database issue):D1100–7. doi:10.1093/nar/gkr777

35. Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics* (2007) **8**:353. doi:10.1186/1471-2105-8-353

36. Tiedemann RE, Zhu YX, Schmidt J, Shi CX, Sereduk C, Yin H, et al. Identification of molecular vulnerabilities in human multiple myeloma cells by RNA interference lethality screening of the druggable genome. *Cancer Res* (2012) **72**(3):757–68. doi:10.1158/0008-5472.CAN-11-2781

37. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* (2011) **39**(Database issue):D214–9. doi:10.1093/nar/gkq1020

38. Chen Y, Zhang Y, Yin Y, Gao G, Li S, Jiang Y, et al. SPD – a web-based secreted protein database. *Nucleic Acids Res* (2005) **33**(Database issue):D169–73. doi:10.1093/nar/gki093

39. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* (2002) **298**(5600):1912–34. doi:10.1126/science.1075762

40. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* (2004) **20**(3):389–98. doi:10.1093/bioinformatics/btg421

41. Wren JD. A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data

divide. *Bioinformatics* (2009) **25**(13):1694–701. doi:10.1093/bioinformatics/btp290

42. Daum JR, Wren JD, Daniel JJ, Sivakumar S, McAvoy JN, Potapova TA, et al. Ska3 is required for spindle checkpoint silencing and the maintenance of chromosome cohesion in mitosis. *Curr Biol* (2009) **19**(17):1467–72. doi:10.1016/j.cub.2009.07.017

43. Lupu C, Zhu H, Popescu NI, Wren JD, Lupu F. Novel protein ADTRP regulates TFPI expression and function in human endothelial cells in normal conditions and in response to androgen. *Blood* (2011) **118**(16):4463–71. doi:10.1182/blood-2011-05-355370

44. Clemmensen SN, Bohr CT, Rørvig S, Glenthøj A, Mora-Jensen H, Cramer EP, et al. Olfactomedin 4 defines a subset of human neutrophils. *J Leukoc Biol* (2012) **91**(3):495–500. doi:10.1189/jlb.0811417

45. Towner RA, Jensen RL, Colman H, Vaillant B, Smith N, Casteel R, et al. ELTD1, a potential new biomarker for gliomas. *Neurosurgery* (2013) **72**(1):77–90. doi:10.1227/NEU.0b013e318276b29d discussion 91.

46. John T, Black MA, Toro TT, Leader D, Gedye CA, Davis ID, et al. Predicting clinical outcome through molecular profiling in stage III melanoma. *Clin Cancer Res* (2008) **14**(16):5173–80. doi:10.1158/1078-0432.CCR-07-4170

47. Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, et al. A gene expression signature associated with survival in metastatic melanoma. *J Transl Med* (2006) **4**:50. doi:10.1186/1479-5876-4-50

48. Journe F, Id Boufker H, Van Kempen L, Galibert MD, Wiedig M, Salès F, et al. TYRP1 mRNA expression in melanoma metastases correlates with clinical outcome. *Br J Cancer* (2011) **105**(11):1726–32. doi:10.1038/bjc.2011.451

49. Timar J, Gyorffy B, Raso E. Gene signature of the metastatic potential of cutaneous melanoma: too much for too little? *Clin Exp Metastasis* (2010) **27**(6):371–87. doi:10.1007/s10585-010-9307-2

50. Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu YL, Adams S, et al. Immune profile and mitotic index of metastatic

melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci U S A* (2009) **106**(48):20429–34. doi:10.1073/pnas.0905139106

51. Jönsson G, Busch C, Knappskog S, Geisler J, Miletic H, Ringnér M, et al. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin Cancer Res* (2010) **16**(13):3356–67. doi:10.1158/1078-0432.CCR-09-2509

52. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* (2003) **19**(2):185–93. doi:10.1093/bioinformatics/19.2.185

53. Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, et al. Integrative analysis of the melanoma transcriptome. *Genome Res* (2010) **20**(4):413–27. doi:10.1101/gr.103697.109

54. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* (2012) **44**(2):133–9. doi:10.1038/ng.1026

55. Palavalli LH, Prickett TD, Wunderlich JR, Wei X, Burrell AS, Porter-Gill P, et al. Analysis of the matrix metalloproteinase family reveals that MMP8 is often mutated in melanoma. *Nat Genet* (2009) **41**(5):518–20. doi:10.1038/ng.340

56. Prickett TD, Agrawal NS, Wei X, Yates KE, Lin JC, Wunderlich JR, et al. Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nat Genet* (2009) **41**(10):1127–32. doi:10.1038/ng.438

57. Prickett TD, Wei X, Cardenas-Navia I, Teer JK, Lin JC, Walia V, et al. Exon capture analysis of G protein-coupled receptors identifies activating mutations in GRM3 in melanoma. *Nat Genet* (2011) **43**(11):1119–26. doi:10.1038/ng.950

58. Stark MS, Woods SL, Gartside MG, Bonazzi VF, Dutton-Regester K, Aoude LG, et al. Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. *Nat Genet* (2012) **44**(2):165–9. doi:10.1038/ng.1041

59. Turajlic S, Furney SJ, Lambros MB, Mitsopoulos C, Kozarewa I, Geyer FC, et al. Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Res* (2012) 22(2):196–207. doi:10.1101/gr.125591.111

60. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* (2011) 43(5):442–6. doi:10.1038/ng.810

61. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012) 483(7391):603–7. doi:10.1038/nature11003

62. Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol* (1995) 57(1):289–300.

63. 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. doi:10.1101/gr.107524.110

64. Sales G, Calura E, Cavalieri D, Romualdi C. graphite – a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* (2012) 13:20. doi:10.1186/1471-2105-13-20

65. Warnes G. Gtools: Various R Programming Tools. *R Package 2.7.0.* (2012).

66. Kelleher FC, McArthur GA. Targeting NRAS in melanoma. *Cancer J* (2012) 18(2):132–6. doi:10.1097/PPO.0b013e31824ba4df

67. Jensen TO, Hoyer M, Schmidt H, Moller HJ, Steiniche T. The prognostic role of protumoral macrophages in AJCC stage I/II melanoma [abstract]. *J Clin Oncol* (2008) 26(15S):9017.

68. Davies MA, Samuels Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene* (2010) 29(41):5545–55. doi:10.1038/onc.2010.323

69. Yancovitz M, Litterman A, Yoon J, Ng E, Shapiro RL, Berman RS, et al. Intra- and inter-tumor heterogeneity of BRAF(V600E) mutations in primary and metastatic melanoma. *PLoS ONE* (2012) 7(1):e29336. doi:10.1371/journal.pone.0029336

70. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* (2003) 31(13):3812–4. doi:10.1093/nar/gkg509

71. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods* (2010) 7(4):248–9. doi:10.1038/nmeth0410-248

72. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* (2012) 7(10):e46688. doi:10.1371/journal.pone.0046688

73. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012) 489(7414):57–74. doi:10.1038/nature11247

74. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* (2012) 22(9):1775–89. doi:10.1101/gr.132159.111

75. Khaitan D, Dinger ME, Mazar J, Crawford J, Smith MA, Mattick JS, et al. The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer Res* (2011) 71(11):3852–62. doi:10.1158/0008-5472.CAN-10-4460

76. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* (2012) 489(7414):109–13. doi:10.1038/nature11279

77. de Graaf CA, van Steensel B. Chromatin organization: form to function. *Curr Opin Genet Dev* (2012) 23(2):185–90. doi:10.1016/j.gde.2012.11.011

78. Sigalotti L, Covre A, Fratta E, Parisi G, Colizzi F, Rizzo A, et al. Epigenetics of human cutaneous melanoma: setting the stage for new therapeutic strategies. *J Transl Med* (2010) 8:56. doi:10.1186/1479-5876-8-56

79. Fratta E, Sigalotti L, Colizzi F, Covre A, Nicolay HJ, Danielli R, et al. Epigenetically regulated clonal heritability of CTA expression profiles in human melanoma. *J Cell Physiol* (2010) 223(2):352–8. doi:10.1002/jcp.22040

80. Howell PM Jr., Liu S, Ren S, Behlen C, Fodstad O, Riker AI. Epigenetics in human melanoma. *Cancer Control* (2009) 16(3):200–18.

81. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* (2011) 12(4):R41. doi:10.1186/gb-2011-12-4-r41

82. Sanchez-Garcia F, Akavia UD, Mozes E, Pe'er D. JISTIC: identification of significant targets in cancer. *BMC Bioinformatics* (2010) 11:189. doi:10.1186/1471-2105-11-189

83. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. *Cell* (2010) 143(6):1005–17. doi:10.1016/j.cell.2010.11.013

84. Flaherty KT, Hodi FS, Fisher DE. From genes to drugs: targeted strategies for melanoma. *Nat Rev Cancer* (2012) 12(5):349–61. doi:10.1038/nrc3218

85. Utikal J, Schadendorf D, Ugurel S. Serologic and immunohistochemical prognostic biomarkers of cutaneous malignancies. *Arch Dermatol Res* (2007) 298(10):469–77.