


METHODOLOGY ARTICLE

Open Access



Whole exome sequencing in the rat

Julie F. Foley^{1*} , Dhiral P. Phadke⁴, Owen Hardy⁵, Sara Hardy⁵, Victor Miller⁵, Anup Madan⁶, Kellie Howard⁷, Kimberly Kruse⁷, Cara Lord⁶, Sreenivasa Ramaiahgari¹, Gregory G. Solomon², Ruchir R. Shah⁴, Arun R. Pandiri³, Ronald A. Herbert³, Robert C. Sills³ and B. Alex Merrick¹

Abstract

Background: The rat genome was sequenced in 2004 with the aim to improve human health altered by disease and environmental influences through gene discovery and animal model validation. Here, we report development and testing of a probe set for whole exome sequencing (WES) to detect sequence variants in exons and UTRs of the rat genome. Using an in-silico approach, we designed probes targeting the rat exome and compared captured mutations in cancer-related genes from four chemically induced rat tumor cell lines (C6, FAT7, DSL-6A/C1, NBTII) to validated cancer genes in the human database, Catalogue of Somatic Mutations in Cancer (COSMIC) as well as normal rat DNA. Paired, fresh frozen (FF) and formalin-fixed, paraffin-embedded (FFPE) liver tissue from naive rats were sequenced to confirm known dbSNP variants and identify any additional variants.

Results: Informatics analysis of available gene annotation from rat RGSC6.0/rn6 RefSeq and Ensembl transcripts provided 223,636 unique exons representing a total of 26,365 unique genes and untranslated regions. Using this annotation and the Rn6 reference genome, an in-silico probe design generated 826,878 probe sequences of which 94.2% were uniquely aligned to the rat genome without mismatches. Further informatics analysis revealed 25,249 genes (95.8%) covered by at least one probe and 23,603 genes (93.5%) had every exon covered by one or more probes. We report high performance metrics from exome sequencing of our probe set and Sanger validation of annotated, highly relevant, cancer gene mutations as cataloged in the human COSMIC database, in addition to several exonic variants in cancer-related genes.

Conclusions: An in-silico probe set was designed to enrich the rat exome from isolated DNA. The platform was tested on rat tumor cell lines and normal FF and FFPE liver tissue. The method effectively captured target exome regions in the test DNA samples with exceptional sensitivity and specificity to obtain reliable sequencing data representing variants that are likely chemically induced somatic mutations. Genomic discovery conducted by means of high throughput WES queries should benefit investigators in discovering rat genomic variants in disease etiology and in furthering human translational research.

Keywords: Whole exome sequencing, Next generation sequencing, C6, FAT7, DSL-6A/C1, NBTII, Sanger, COSMIC

Background

The laboratory rat is a useful mammalian model for the translation and validation of human gene-function discovery toward understanding the interplay between genetics, environmental influences and disease biology. As an experimental animal in toxicology and safety pharmacology applications, the rat is often the model of choice because of its relatively large size and its biologic relevance to

human physiology, disease and histopathology. In the past decade, its research popularity has continued because of molecular sequencing advancements through Next Generation Sequencing (NGS). The rat genome was sequenced in 2004 [1] and refinements such as an RNA-Seq expression atlas [2] and genomic updates continue to improve our understanding [3] of the species. Many rat strains are broadly used in toxicology and pharmacology studies and recent developments in genome editing technologies such as CRISPR/Cas-9, have significantly increased the library of available rat strains as targeted disease models for gene discovery and validation [4, 5].

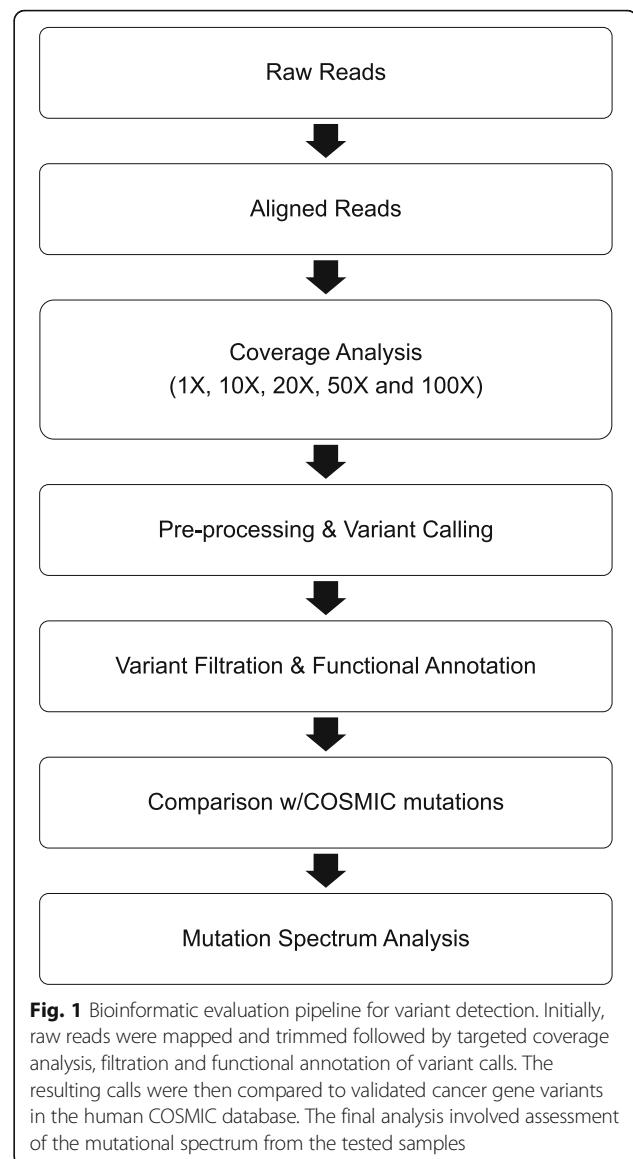
* Correspondence: foley1@niehs.nih.gov

¹Biomolecular Screening Branch, National Institute of Environmental Health Sciences, 111 T.W. Alexander Dr. Research Triangle Park, Durham, NC, USA
Full list of author information is available at the end of the article



Exome sequencing provides an efficient method to examine sequence variants in coding regions that are related to disease without sequencing the entire genome [6]. Conventional sequencing methods rely upon hybridization of probes from fragmented DNA that are designed around sequences of exonic regions. Various platforms have adopted different probe types and capture chemistries involving overlapping, tiling or gapped probes. Exon size, GC content, repeat elements, and segmental duplications are factors in probe design affecting exon coverage [7]. A recent report on a rat probe set called TargetEC [8], based on the rat reference genome from assembly Rnor_5 was designed to capture coding exonic regions and conserved non-coding regulatory sequences from 13 vertebrate species. The entire probe set covered a 146.8 Mb genomic region. Two control inbred rat strains (WTC/Kyo, PVG/Seac) and 2 mutant strains (WTC-swh/Kyo, KFRS4/Kyo) with known disease mutations were examined. The WES performance metrics for capture specificity and sensitivity were acceptable with 85–94% reads from exomes after removing duplicates and ~79% reads on target with an average target depth ranging from 107 to 125-fold. Sequencing validation was minimal, evaluating only two captured mutations previously identified as responsible for disease phenotypes in mutant rat strains [8]. The large target size of capture probes at 146.8 Mb may limit the utility of this platform due to the sequence depth required for the detection of low frequency or rare sequence variants. Thus, other novel approaches are needed to achieve sufficient performance that may enable evaluation of potentially human-relevant disease causing mutations in the rat genome.

Here we describe an in-silico probe design for the evaluation and validation of a WES platform specifically for the rat based on the updated Rnor_6.0 assembly using RefSeq and Ensembl annotations. Tiling probes with a 1 bp overlap were designed, and libraries constructed, using the Agilent SureSelect® XT target enrichment system. Naive, paired FF and FFPE rat liver DNA samples served as control rat exomes and were compared against four chemically induced, rat tumor cell lines (C6, FAT7, DSL-6A/C1 and NBTII) available through American Type Culture Collection (ATCC). Bioinformatic evaluation (Fig. 1) of the FF-FFPE samples revealed cataloged SNPs as well as some not reported in dbSNP (Build 149; November 7, 2016). We were able to affirm in the chemically induced, tumor cell lines, high quality sequence variants associated with common, cancer-related genes annotated in the COSMIC database (v82). Together these data demonstrate WES can be a valuable tool for high throughput sequencing (HTS) in rat models of disease and chemical exposure.



Methods

Tissue samples

Samples for evaluating the rat exome-seq platform consisted of DNA extracted from control, non-treated rat liver tissue and chemically induced rat tumor cell lines (American Type Culture Collection, Manassas, VA) (Table 1).

Biological replicates of liver from four naive, male, Sprague Dawley retired breeders (CRL:SD, Taconic Biosciences, Hudson, NY), 6–9 months of age, were used as normal controls for the study. Fresh frozen tissue samples served as the “gold standard” for the rat WES evaluation to compare with paired FFPE tissue and the chemically induced, tumor cell lines with documented mutations selected as positive controls for obtaining exonic mutations in cancer-related genes.

Table 1 Rat exome-seq platform study samples

| Sample Name | Type | Treatment | Rat Strain |
|-------------|--------------------------------------|---------------------------------------|----------------|
| C6 | Glioma | N,N-nitroso-methylurea | Wistar |
| FAT7 | Nasal cavity squamous cell carcinoma | Formaldehyde | Fisher-344 |
| DSL-6A/C1 | Pancreatic acinar carcinoma | Azaserine | Lewis |
| NBTII | Surface epithelial bladder carcinoma | N-butyl-N-(4-hydroxybutyl)nitrosamine | Wistar |
| FF1 | Fresh frozen | Normal liver | Sprague Dawley |
| FF2 | Fresh frozen | Normal liver | Sprague Dawley |
| FF3 | Fresh frozen | Normal liver | Sprague Dawley |
| FF4 | Fresh frozen | Normal liver | Sprague Dawley |
| FFPE1 | Formalin-fixed, paraffin-embedded | Normal liver | Sprague Dawley |
| FFPE2 | Formalin-fixed, paraffin-embedded | Normal liver | Sprague Dawley |
| FFPE3 | Formalin-fixed, paraffin-embedded | Normal liver | Sprague Dawley |
| FFPE4 | Formalin-fixed, paraffin-embedded | Normal liver | Sprague Dawley |

Experiments were performed according to the guidelines established in the NIH Guide for the Care and Use of Laboratory Animals (National Research Council, 2011). All animals were treated humanely for alleviation of potential suffering, as approved by the National Institute of Environmental Health Sciences Animal Care and Use Committee.

Following euthanasia with CO₂, paired FF and FFPE samples were prepared from the left lobe of each animal. A single, representative section (3 mm) was fixed overnight (18–24 h) in 10% neutral-buffered formalin, routinely processed and embedded in paraffin. The remainder of the lobe was cubed (3–5 mm), flash-frozen in liquid nitrogen, and stored at –80 °C. FFPE blocks were stored at room temperature and sectioned within two months of tissue embedding. All sectioning was conducted under sterile, nuclease-free conditions. Prior to sectioning, the block was trimmed to minimize paraffin surrounding the tissue. A nuclease-free, water pre-soak for 30 s at room temperature prevented tissue chattering. Three to five, 10 μm sections of the FFPE block, dependent upon tissue area were collected in a sterile cryovial and stored at –20 °C until DNA isolation.

Cell line samples

To meet the study objective of identifying variants in a newly developed rat exome method, we purposely aimed for toxicology relevant, chemically induced cancer cell lines that contained mutations associated with a cancer phenotype. Four, rat tumor cell lines induced by chemical exposure (C6; glioma, FAT7; nasal squamous cell carcinoma, DSL-6A/C1; pancreatic acinar carcinoma and NBTII; surface epithelial bladder tumor) were used for the testing of the WES platform. The number of cells processed for each respective cell line was C6 (3×10^6), FAT7 (3.1×10^6),

DSL-6A/C1 (3.1×10^6), and NBTII (1.3×10^6). Cell lines were thawed, suspended in 10 mL of media according to the manufacturer's specifications and spun (Eppendorf 5810R centrifuge, Hauppauge, NY) at 172 x g (1000 rpm) for 5 min. The supernatant was discarded, the cells suspended in 200 μL of 1X PBS and immediately processed for DNA isolation.

Genomic DNA purification

Qiagen kits (Qiagen, Germantown, MD) were used according to manufacturer's instructions for genomic DNA isolation from the paired FF (QIAamp Fast DNA Tissue Kit) and FFPE (GeneRead DNA FFPE Kit) liver tissue, and the rat tumor cell lines (Blood and Cell Culture DNA Mini Kit). Sample integrity and yield were assessed by the Nanodrop® (Thermo Fisher Scientific, Madison, WI), the Qubit® Fluorometer (Thermo Fisher Scientific) and the Agilent TapeStation® (Santa Clara, CA). Purified DNA samples were stored at –20 °C.

In-silico probe design

RefSeq and Ensembl gene annotations for the rat reference genome from assembly Rnor_6.0 (RGSC, 2014) were downloaded from the UCSC Genome Browser [9]. RefSeq and Ensembl annotations covered protein coding, lncRNAs and miRNAs with no RefSeq overlap and for which exons were removed that were completely contained in other longer exons. Merging the two annotations, the in-silico probe design covered a total of 223,636 exons (26,365 genes) of the reference rat exome. Based on the input targets, 120-base RNA probes were created iteratively by tiling at approximately 1X density (end-to-end) along the sense strand of the reference genome, Rnor_6.0 assembly. Probe filtering was based on the uniqueness criteria of an exon read having a single chromosome location within the Rn6 genome. Low complexity or repetitive probes not meeting the Rn6

genome criteria were removed from the pool. In total, 826,878 probe sequences were combined and used to manufacture a single biotinylated RNA library for target capture.

Exome enrichment and sequencing

Sequencing libraries were created using the Agilent SureSelect^{XT} method with on-bead modifications. Genomic DNA (150 ng) was fragmented to approximately 150 bp with a Covaris sonicator (Woburn, MA) and purified with Agencourt AMPure XP beads (Beckman Coulter Genomics, Brea, CA) according to the manufacturer's specifications. Post-capture library amplifications and quality assessments were also performed according to manufacturer's specifications. Illumina sequencing on the HiSeq2500[®] platform (San Diego, CA) was performed by Q² Solutions (Morrisville, NC) at a 75X exome coverage from 2 × 100 bp paired-end reads. Libraries were bar-coded and multiplexed in a pool of 4 samples over 3 lanes. The Illumina Casava[®] software (v1.8) was used to make base calls. Sequences were output in FASTQ format. Raw data were deposited in GEO (PRJNA434726).

Variant detection

FASTQ files were subjected to quality control with the FastQC tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read pairs were mapped to the Rn6 genome using the BWA alignment tool (v0.7.12) [10]. Any exon reads whose two ends mapped on different chromosomes were discarded and considered ambiguous as the read did not match the probe filtering criteria of an exon matching to a single chromosome. Duplicates were trimmed from the reads using the MarkDuplicates program from Picard tools (v1.99). Utilities from the BED-Tools package were used to obtain the coverage at each target base. Coverage was summarized at the following different levels: 1X, 10X, 20×, 30X and 50X. Variants were called using Genome Analysis Toolkit (GATK; v3.7) [11]. SNVs and INDEL calling was performed with the GATK utility, HaplotypeCaller (https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_haplotypecaller_-_HaplotypeCaller.php). Exonic variant filtration was done with GATK, followed by functional annotation of the variants with the SnpEff tool [12]. To determine if there was human disease relevance between our chemically induced tumor mutations and previously identified mutations found in human malignant tumors, we matched the variants in each cell line to validated mutations present in the COSMIC database with the same amino acid substitutions and/or location as that found in our exome sequence data. The mutation spectrum of the final variant set was analyzed with the R package,

SomaticSignatures (v3.6; Bioconductor), [13] and compared with different mutational processes that generate unique combinations of mutation types termed mutational signatures in the Catalogues of Somatic Mutations In Cancer (COSMIC) database [14].

Sequence validation

SNPs from the four chemically induced, rat tumor cell lines detected by WES with medium (non-synonymous missense, nonsense) or high (frameshift mutations and INDELS) functional effects were selected for orthogonal testing by Sanger Sequencing. Variants included sequencing of three well-studied, cancer-related genes (*Tp53*, *Pik3ca*, and *Ncor1*) along with a subset of 15 additional cancer-related genes (*Nf1*, *Mki67*, *Mllt4*, *Fgfr2*, *Ctnnb1*, *Fat1*, *Clp1*, *Fat4*, *Arid1a*, *Nat1*, *Nat2*, *Setd2*, *Impg1*, *Nbas*, and *Npat*). Primers were designed from 500 bp flanking nucleotide sequences of the Rnor_6 rat assembly for each sequence variant (Additional file 1: Table S1). Two gene deletions (*Cdkn2a*, *Cdkn2b*) called by WES were evaluated by qPCR analysis.

Sanger sequencing

PCR amplification of various regions from the designated cancer-related genes was performed using KAPA Hyper polymerase (Roche Holding AG, Basel, CH) with the primer pairs found in Additional file 2: Table S2. Forward primers were tagged with M13 forward sequences and reverse primers tagged with M13 reverse sequences. The reaction was as follows: after preheating at 95 °C for 5 min, amplification consisted of 30 cycles at 98 °C for 20 s, 65 °C for 15 s and 72 °C for 15 s, and a final extension at 72 °C for 5 min. PCR products were purified using the AMPure[®] XP (Beckmann Coulter, Brea, CA) and quantified using the NanoDrop[®]. DNA (10 ng) was sequenced using the BigDye Terminator[®] cycle sequencing kit (v3.1; Thermo Fisher Scientific). Following purification using the Centri-Sep[®] Spin Columns (Thermo Fisher Scientific), the nucleotide sequences were determined using an ABI3730XL Genetic Analyzer (Thermo Fisher Scientific). Sequence data was aligned using Sequencher[®] DNA sequence analysis software (v5.2.4; Gene Codes Corporation, Ann Arbor, MI).

qPCR analysis for gene (*Cdkn2a* and *Ckdn2b*) deletion

DNA study samples from the four rat cell lines along with one normal rat liver control sample and two no template controls were subjected to qPCR analysis using TaqMan[®] qPCR CNV Assays on the Applied Biosystems 7900HT Fast Real-Time PCR System (Foster City, CA). TaqMan[®] CNV Master mix was prepared using 1.25 µL of 20× from either one of three custom designed assays specific to *Cdkn2a* or *Cdkn2b* and one of the two copy number reference assays (*APPRKTP* and *APRWFDm*).

Reactions were set using 2 μ L (5 ng/ μ L) of DNA samples and 8 μ L of CNV master mix (Thermo Fisher Scientific). Each sample was analyzed in duplicate with denaturation at 95 °C for 10 min, followed by 40 cycles at 95 °C for 15 s, and 60 °C for 1 min. Data were analyzed using Applied Biosystems CopyCaller™ Software (v2.1; Thermo Fisher Scientific).

Results

Rat exome capture probes were designed to capture 71 Mb of the rat genome. An initial experiment was performed to test the capture ability of the probe set on exons and flanking regions from 26,365 rat genes. Enriched DNA fragments were sequenced on an Illumina HiSeq2500® platform from four paired FF and FFPE rat liver samples and four chemically induced, rat tumor cell lines.

Probe performance

Sequencing was performed at an average of 75-fold depth in exome coverage. The mean \pm SEM number of bases sequenced was 19.1 \pm 0.202 Gb for the FF liver, 17.3 \pm 0.335 Gb for the FFPE liver and 18.2 \pm 0.297 Gb for the cell lines with minimal variability among sample types. Liver and cell line sample reads mapped at 99% to the Rn6 reference genome (Table 2).

Approximately 80% of the reads aligned to target exons in FF and cell line samples. Target read duplicates were relatively low at a mean of 23% \pm 0.6 for FF liver and 29% \pm 1.0 for cell lines, but were almost doubled in FFPE samples at 45% \pm 1.3.

Rat exome-Seq performance

Performance characteristics include the number of reads per base for each exon and uniformity for depth of

coverage. Minimum depth of coverage is typically 10–20-fold for accurate base calls [15, 16]. We analyzed base pair coverage at 1X, 10X, 20X, 30X and 50X to examine the effects of increasing levels of coverage stringency (Fig. 2). Capture sensitivity was consistent across the three sample types and within the sample groups. At a minimum coverage depth of 10X, ~98% of the target bases were sequenced for all samples with the exception of one FFPE sample (FFPE3), which may be related to higher sample fragmentation with formalin fixation [17]. A more stringent depth of coverage cut off at 50X was selected to minimize false positive variant detection; 50X coverage showed acceptable base pair coverage for the FF (88%) and cell line (84%) samples, while we noted the FFPE coverage lowered to ~57%. Target regions not covered by any reads were negligible with less than 1% of non-covered bases.

We also looked at the uniformity of coverage for up to 500 bp reads for the cell lines and paired FF-FFPE samples and found the depth of coverage consistent across samples within the cell lines, FF and FFPE groups. Cell line depth of coverage distribution peaked at ~80 reads per ~500,000 bases (Fig. 3a-d). The distribution was broader for the FF samples with the peak at about ~80–100 reads with 400,000 bases (Fig. 3e-h). For the FFPE samples, the peak of reads shifted to the left with decreased reads (20–40) covering 600,000–800,000 bases (Fig. 3i-l). FFPE3 was an outlier with 20–40 reads covering over a million bases, again consistent with greater DNA fragmentation. Depth of coverage plot distributions were consistent with previously published work [15].

We examined the distribution of gene reads for bias by plotting all the sample exonic variants and comparing them to the distribution of RefSeq genes across the rat chromosomes (Additional file 3: Figure S1). With the

Table 2 Summary statistics for rat WES reads

| Sample | Total Reads | Aligned Reads | Aligned Reads (%) | Reads in Target Exons | Reads on Target (%) | Duplicate Reads (%) |
|-----------|-------------|---------------|-------------------|-----------------------|---------------------|---------------------|
| C6 | 171M | 170 M | 98.9 | 135 M | 79.5 | 32.3 |
| DSL-6A/C1 | 179 M | 178 M | 99.1 | 141 M | 79.1 | 27.5 |
| FAT7 | 185 M | 183 M | 99.1 | 146 M | 79.6 | 29.6 |
| NBTII | 176 M | 174 M | 99.1 | 139 M | 79.9 | 27.6 |
| FF1 | 187 M | 185 M | 99.2 | 146 M | 79.0 | 22.7 |
| FF2 | 181 M | 180 M | 99.1 | 141 M | 78.4 | 21.8 |
| FF3 | 189 M | 187 M | 99.0 | 150 M | 79.8 | 24.9 |
| FF4 | 189 M | 187 M | 99.0 | 149 M | 79.9 | 23.0 |
| FFPE1 | 172 M | 159 M | 92.0 | 118 M | 74.4 | 46.9 |
| FFPE2 | 174 M | 165 M | 95.0 | 128 M | 77.3 | 42.7 |
| FFPE3 | 160 M | 134 M | 84.0 | 81 M | 60.3 | 49.1 |
| FFPE4 | 170 M | 150 M | 88.2 | 105 M | 70.4 | 44.1 |

Read length was 101 bp for all samples

M Million

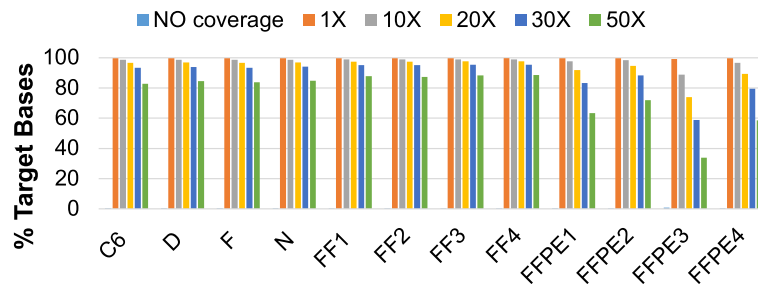


Fig. 2 Breadth of reference genome coverage. The percentage of target bases covering the rat Rn6 reference genome is shown at 1X, 10X, 20X, 30X and 50X depth of coverage. Rigorous testing at 50X demonstrated strong coverage of the rat reference genome by the sequenced fragments

exception of chromosome 20, variant reads were uniformly distributed across all chromosomes. With regards to chromosome 20, there were several variants related to the RT1 family genes (histocompatibility complex family), perhaps indicating a data rich region accounting for the higher distribution of reads than the other chromosomes.

Variant identification

Variant identification was a multistep process where a coverage depth of 50X and an alternate allele depth of 20X were considered discriminately stringent for sensitivity and accuracy. Variants were called using GATK v3.7 program. SNVs and INDELs were identified with the GATK v3.7 utility, HaplotypeCaller[®]. Exonic variants

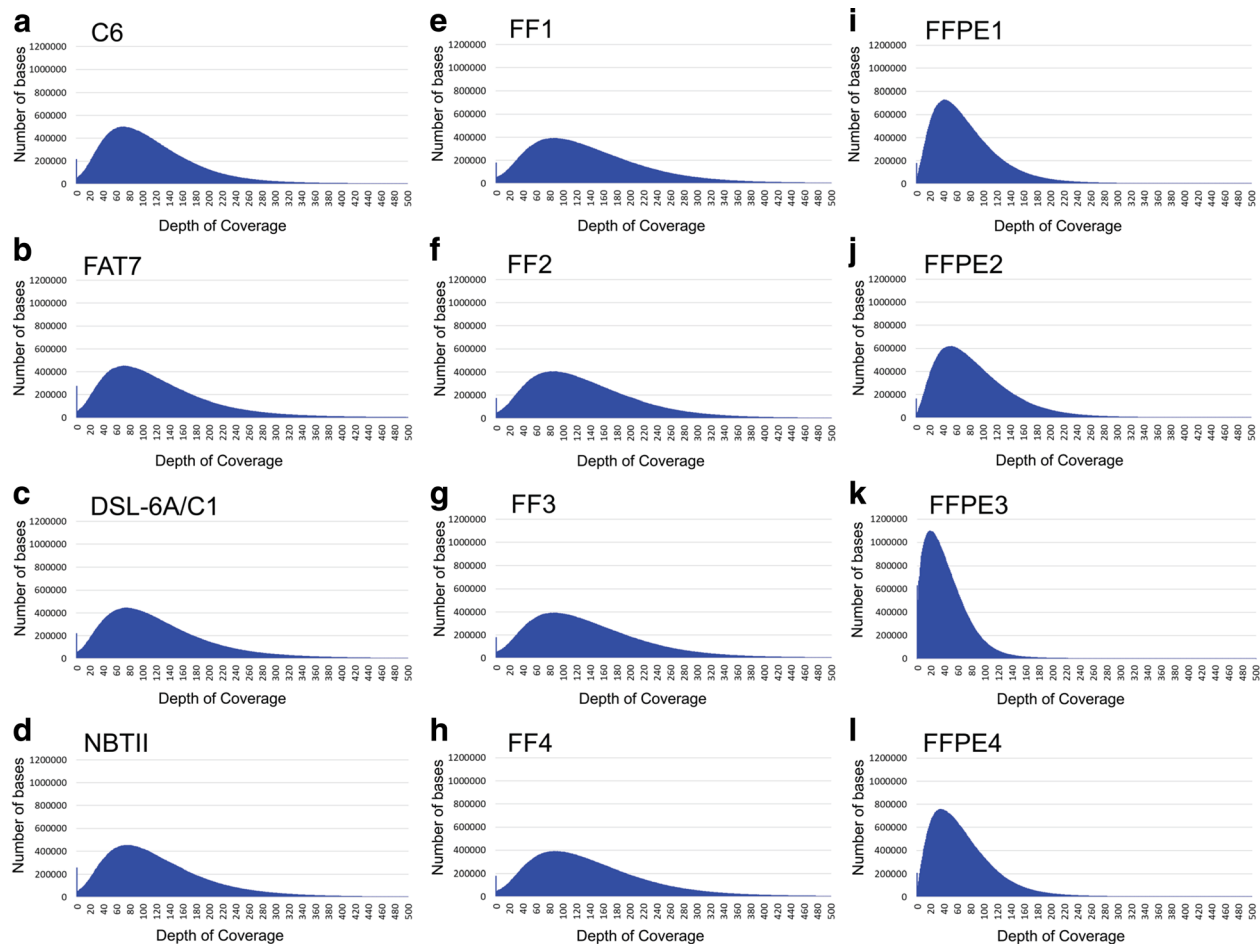


Fig. 3 Uniformity of coverage for up to 500 bp reads for the cell lines and paired FF-FFPE samples. **a-d** Depth of coverage distribution for the C6, FAT7, DSL-6A/C1 and NBTII cell lines. **e-h** Depth of coverage distribution for the fresh frozen (FF) liver tissue. **i-l** Depth of coverage distribution for the formalin-fixed, paraffin-embedded (FFPE) liver tissue

were filtered with GATK, followed by functional annotation of the variants with the SnpEff tool. Intronic and intergenic variants were removed leaving only coding regions and UTR variants (Table 3).

A total of 153,369 SNPs were identified across all exonic variants in the 50 Mb exome component of the design with 85,036 registered in dbSNP (Build 149; November 7, 2016), leaving 68,333 unique SNPs. We used FF samples as the “gold standard” so that 22,685 total SNPs were identified with 14,240 annotated or registered in dbSNP and 8445 identified as non-annotated, not registered in dbSNP. Compared to the FF tissue samples, FFPE variants were substantially decreased (15,944 total, 10,193 annotated and 5751 non-annotated).

Rat tumor cell lines were selected as a proof of principle model to test the rat WES platform, since variants should be plentiful and diverse in chemically induced, tumor derived cancer cell lines. After filtering annotated and non-annotated FF-FFPE normal tissue variants from the tumor cell line data, numerous annotated and non-annotated variants based on dbSNP were identified in each cell line (Table 4).

NBTII, the bladder cell line contained the most variants (10,123), followed by the glioma C6 line (5387) the FAT7, formaldehyde line (3551) and the DSL-6A/C1, pancreatic cancer line (3277). To determine if there was human disease relevance between our chemically induced tumor mutations and previously identified mutations found in human malignant tumors, we matched the variants in each cell line to validated mutations present in the COSMIC database with the same amino acid substitutions and/or location as that found in our exome sequence data. We report for each tumor cell line, relevant cancer-related genes and those genes associated with a specific cancer type.

The C6 rat glioma tumor cell line, chemically induced by N,N-nitrosomethylurea in an outbred Wistar rat [18] and morphologically similar to glioblastoma multiforme, is an important human translational research model for malignant glial neoplasms [19, 20]. Exome sequencing of the C6 glioma captured 3864 non-annotated variants of which 3447 were designated as SNPs and 417 as INDELS. Genes of interest previously reported with

human relevance to glioblastomas include *Cdkn2a*, *Cdkn2b*, *Tp53* and *Pik3ca*. Our results verify these reports for all these genes [20, 21]. Deletion of *Cdkn2a* and *Cdkn2b* and the absence of any *Tp53* mutations as detected by exome sequencing is in agreement with the mutant *p16/CDK2a/Ink4a* locus [20, 21] with no expression of p16 and p19ARF mRNAs and a wildtype p53 previously reported for the C6 line [21, 22]. The oncogene, *Pik3ca*, matched the COSMIC database with an identical amino acid substitution (C90Y) and location as reported in Grade IV astrocytomas [14].

Formaldehyde, an abundantly produced chemical classified as a rodent and human carcinogen [23] has undergone extensive genotoxic, carcinogenic and teratogenic studies. Early mechanistic studies focused on *Tp53* changes, specifically a point mutation (cGt – cAt) in a well conserved region at codon 271 (R271H) [24–26]. We confirmed this point mutation using the FAT7 cell line, a formaldehyde induced squamous cell carcinoma of the nasal cavity derived from a Fisher-344 rat. This mutation was not documented in any of the other exome sequenced tumor cell lines. We identified 2340 non-annotated variants of which 1963 were designated as SNPs and 377 as INDELS. Other reported molecular endpoints characterizing formaldehyde exposure are induced missense mutations mainly at G:C base pairs [27, 28]. We checked for this mutation and found 4–5% G:C and 4–5% C:G substitutions, accounting for ~10% of the base substitutions in the FAT7 cell line. The majority of mutations in this line are from C:T (~20%) and G:A (~20%). It is important to note these results are not specific to the formaldehyde-exposed cell line, as they are seen in each of the evaluated cell lines. Exome sequencing can help distinguish true molecular changes secondary to formaldehyde exposure and those specific to squamous cell carcinomas lending additional insight into mechanisms of formaldehyde-induced carcinoma, which could hold significant relevance for environmental or occupational exposure studies.

DSL-6A/C1 is a pancreatic ductal carcinoma cell line derived from a transplantable DSL-6 acinar cell carcinoma of an azaserine-treated Lewis rat [29]. Exome sequencing identified 2208 non-annotated variants, of

Table 3 All exonic variants detected in the rat exome-seq samples

| Sample | Total Number of SNPs (%) | Annotated SNPs (%) (based on dbSNP) | Non-annotated SNPs (%) (based on dbSNP) |
|-----------|--------------------------|-------------------------------------|-----------------------------------------|
| C6 | 30,529 | 15,945 (52.2%) | 14,584 (47.8%) |
| FAT7 | 24,167 | 13,818 (57.2%) | 10,349 (42.8%) |
| DSL-6A/C1 | 22,060 | 12,719 (57.7%) | 9341 (42.3%) |
| NBTII | 37,984 | 18,121 (47.7%) | 19,863 (52.3%) |
| FF | 22,685 | 14,240 (62.7%) | 8445 (37.2%) |
| FFPE | 15,944 | 10,193 (63.9%) | 5751 (36.1%) |

Variant detection at 50X depth of coverage and alternate allele depth of 20x

Table 4 Cell line specific exonic variants detected in the rat exome-seq samples

| Sample | Total Number of Variants (SNPs + INDELS) | Annotated Variants (%) (based on dbSNP) | Non-annotated Variants (%) (based on dbSNP) |
|-----------|------------------------------------------|-----------------------------------------|---------------------------------------------|
| C6 | 5387 | 1523 (28.3%) | 3864 (71.7%) |
| FAT7 | 3551 | 1211 (34.1%) | 2340 (65.9%) |
| DSL-6A/C1 | 3277 | 1069 (32.6%) | 2208 (67.4%) |
| NBTII | 10,123 | 2430 (24.0%) | 7693 (76.0%) |

Variant detection at 50X depth of coverage and alternate allele depth of 20x

which 1815 were SNPs and 393 were INDELS. A mutation unique to DSL-6A/C1 is *Cttnb1*, a reported gene in 20–25% of human acinar cell carcinomas [30]. Our exome-seq data for this mutation matched the COSMIC database location and amino acid substitution. As in the C6 glioma line, there were homozygous deletions of *Cdkn2a* and neighboring *Cdkn2b*. This alteration reportedly occurs more frequently in human pancreatic ductal carcinomas [31]. *Tp53* lacked any genetic alterations. Rat exome-seq findings for DSL-6A/C1 were concordant with human pancreatic acinar cell carcinoma genetic alterations. Acinar cell carcinomas lack frequent alterations in genes commonly mutated in pancreatic ductal adenocarcinoma such as *KRAS* and *Tp53* [30, 31]. Although the ATCC specifies the DSL-6A/C1 line as a pancreatic ductal cell carcinoma from a transplanted acinar cell carcinoma, the ductal phenotype is not adopted until after 2–3 weeks of cultured growth when the cells lose amylase production [29]. Since we processed the cells directly after thawing and did not culture, it appears the acinar cells possess both acinar (*Cttnb1*, wild type *Tp53*) and ductal (deletion of *Cdkn2a* and *Cdkn2b*) traits related to each pathologic phenotype.

The total number of variants (10,123) in the NBTII bladder carcinoma cell line induced in a Wistar rat treated with N-butyl-N-(4-hydroxybutyl)nitrosamine far exceeded the total number of variants called in the other three cell lines (Table 4). Non-annotated SNPs (7693), and INDELS (828) were identified. Two *Tp53* point mutations present in codons 211 (R211W) and 230 (I230T) had the same amino acid substitutions and location found in the COSMIC database. Missense mutations were recognized in *Ncor1* and *Pik3ca*, oncogenes of central importance to human cancer as known driver mutations. Other exonic variants identified by the rat WES platform with relevance in human bladder carcinoma include *RBI*, *Muc16*, *Muc5b*, *NAT1*, *NAT2*, *Stag2*, *Stag3* and *Arid1a*. All these genes were unique to the NBTII bladder cell line.

Sequence validation

Sanger sequencing provided an orthogonal means of testing data accuracy. We analyzed a select set of called variants of moderate (non-synonymous, missense,

nonsense) to high (frameshift mutations and INDELS) impact in cancer-related genes with exonic variants unique to a single cell line or found across multiple lines. Furthermore, to add richness to our validation assessment, we included cancer genes with mutations matching amino acid substitutions and/or locations in the COSMIC database. The first gene validation set consisted of highly relevant “common” cancer genes (*Tp53*, *Ncor1*, *Pik3ca*, *Cdkn2a* and *Cdkn2b*) shared by many tumors. A second set consisted of an additional 15 cancer-related genes (*Nf1*, *Mki67*, *Mllt4*, *Fgfr2*, *Cttnb1*, *Fat1*, *Clp1*, *Fat4*, *Arid1a*, *Nat1*, *Nat2*, *Setd2*, *Impg1*, *Nbas* and *Npat*) (Table 5). We confirmed all exome-seq variants in the “common” gene set in all four tumor cell lines; the presence of *Tp53* mutations in the FAT7 (R271H) and NBTII lines (R211W, I230T), and its absence in the C6 and DSL-6A/C1 cell lines, missense mutations for the oncogenes *Ncor1* (E544D) across all four cell lines and *Pik3ca* in the C6 (C90Y) and the NBTII (M811 T) cell lines. All mutations in the ‘common’ gene set matched the variant calls in the human COSMIC database. Homozygous deletions of *Cdkn2a* and *Cdkn2b* were confirmed by qPCR analysis in cell lines C6 and DSL-6A/C. There was complete concordance between the rat WES and Sanger sequencing methods for the common gene set.

Next, we confirmed genotypes from cancer-related genes assigned to the second validation set. We identified 38 variants in 15 genes of which only four were inaccurately called. These miscalls were possibly due to low coverage or low allelic frequency. Associated with 12 of the genes were 26 variants where the amino acid substitution and/or location matched the validated mutations in COSMIC database. False negative calls (×4) found by Sanger were identified in a single gene, *Nbas*. The variant calls (T837 N and A841V) were correctly called in the C6 and DSL-6A/C1 tumor lines; however, these mutations were missed by WES in the FAT7 and NBTII lines.

Mutation Spectrum

We looked at all exonic variants with respect to the rat reference genome, Rn6, and analyzed the mutational spectrum using the Kullback-Leibler divergence to

Table 5 Variant candidate validation by Sanger sequencing

| Chromosome #: Position | Gene | Exon | Codon Change | C6 | | | FAT7 | | | DSL 6A/C1 | | | NBTH1 | | |
|---------------------------|---------------------------------|------------|----------------------------|-------------------------|-------------|---------------------------|-------------------------|------------------------|---------------------------|-------------------------|-------------------------|---------------------------|------------------------|------------------------|--------------|
| | | | | Coverage | | AF | Coverage | | AF | Coverage | | AF | Coverage | | AF |
| | | | | Total | Allele | | Total | Allele | | Total | Allele | | Total | Allele | |
| Chr10:56196111 | <i>Tp53</i> ^a | R271H | cGt/cAt | WT | WT | - | 71^d | 71 | 1.000 | WT | WT | - | WT | WT | - |
| Chr10:56195619 | <i>Tp53</i> | R211W | Cgg/Tgg | WT | WT | - | WT | WT | - | WT | WT | - | 197^e | 113 | 0.574 |
| Chr10:56195677 | <i>Tp53</i> | I230T | aTc/aCc | WT | WT | - | WT | WT | - | WT | WT | - | 174 | 90 | 0.529 |
| Chr10:48699815 | <i>Ncor1</i> | E544D | gaG/gaC | 194 | 85 | 0.438 | 157 | 157 | 1.000 | 166 | 166 | 1.000 | 191 | 191 | 1.000 |
| Chr2:118851700 | <i>Pik3ca</i> ^b | M811 T | aTg/aCg | WT | WT | - | WT | WT | - | WT | WT | - | 227 | 111 | 0.489 |
| Chr2:118831618 | <i>Pik3ca</i> | C90Y | tGt/tAt | 129 | 62 | 0.481 | WT | WT | - | WT | WT | - | WT | WT | - |
| Chr5 | Cdkn2a | Deletion | | WT | WT | - | WT | WT | - | WT | WT | - | WT | WT | - |
| Chr5 | Cdkn2b | Deletion | | WT | WT | - | WT | WT | - | WT | WT | - | WT | WT | - |
| Chr10:66790063 | Nf1 | Q962* | Cag/Tag | 151 | 79 | 0.523 | WT | WT | - | WT | WT | - | WT | WT | - |
| Chr1:208011800 | Mki67 | G207R | Ggg/Agg | 314/_f | 68/- | 0.217/_f | WT | WT | - | 442/_f | 202/_f | 0.457/_f | WT | WT | - |
| Chr1:53731379 | Mllt4 | Q1440* | Caa/Taa | WT | WT | - | 111/_f | 30/- | 0.270/_f | WT | WT | - | WT | WT | - |
| Chr1:200672370 | Fgfr2 | V70 M | Gtg/Atg | 126 | 126 | 1.000 | 159 | 159 | 1.000 | WT | WT | - | WT | WT | - |
| Chr8:129618253 | <i>Ctnnb1</i> | D32V | gAt/gTt | WT | WT | - | WT | WT | - | 55 | 20 | 0.364 | WT | WT | - |
| Chr16:50398579 | Fat1 | L2464P | cTc/cCc | WT | WT | - | WT | WT | - | WT | WT | - | 154 | 62 | 0.403 |
| Chr16:50485429 | Fat1 | M536 T | aTg/aCg | 131 | 67 | 0.511 | 149 | 149 | 1.000 | WT | WT | - | 196 | 98 | 0.500 |
| Chr16:50485954 | Fat1 | F361S | tTc/tCc | WT | WT | - | 175 | 175 | 1.000 | WT | WT | - | WT | WT | - |
| Chr3:72125722 | <i>Clp1</i> ^c | M417 | atg/ | WT | WT | - | WT | WT | - | 165/_f | 25/- | 0.152/_f | WT | WT | - |
| Chr2:125846677 | <i>Fat4</i> | I3077V | Att/Gtt | WT | WT | - | WT | WT | - | 200 | 200 | 1.000 | WT | WT | - |
| Chr2:125754029 | <i>Fat4</i> | S627 T | aGt/aCt | WT | WT | - | WT | WT | - | WT | WT | - | 94 | 37 | 0.394 |
| Chr2:125754184 | <i>Fat4</i> | L679F | Ctc/Tcc | WT | WT | - | WT | WT | - | WT | WT | - | 143 | 84 | 0.587 |
| Chr5:151918885 | <i>Arid1a</i> | Y658* | taT/taA | WT | WT | - | WT | WT | - | WT | WT | - | 78 | 29 | 0.372 |
| Chr16:23972323 | Nat1 | S15 L | tCa/tTa | WT | WT | - | WT | WT | - | WT | WT | - | 300 | 154 | 0.513 |
| Chr16:23961976 | <i>Nat2</i> | L52* | tTa/tAa | WT | WT | - | WT | WT | - | WT | WT | - | 197 | 90 | 0.457 |
| Chr8:118824049 | <i>Setd2</i> | R792TEPSVR | agg/aCTGAACCTT CAGTTAgg | WT | WT | - | 95 | 95 | 1.000 | 103 | 103 | 1.000 | WT | WT | - |
| Chr8:118824049 | <i>Setd2</i> | R79TESSVR | agg/aCTGAATCTT CAGTTAgg | WT | WT | - | WT | WT | - | WT | WT | - | 120 | 38 | 0.317 |
| Chr8:87845140 | <i>Impg1</i> | T133I | aCc/aTc | 151 | 74 | 0.490 | WT | WT | - | WT | WT | - | WT | WT | - |
| Chr6:38567303 | <i>Nbas</i> | T837 N | aCc/aAc | 56 | 56 | 1.000 | WT/_f | WT/_f | - | 60 | 60 | 1.000 | WT/_f | WT/_f | - |
| Chr6:38567315 | <i>Nbas</i> | A841V | gCg/gTg | 53 | 53 | 1.000 | WT/_f | WT/_f | - | 58 | 58 | 1.000 | WT/_f | WT/_f | - |
| Chr8:58154526 | <i>Npat</i> | K1186R | aAg/aGg | 68 | 68 | 1.000 | 50 | 50 | 1.000 | WT | WT | - | WT | WT | - |

^aIn the gene column, bold and italics indicates the variant found in the rat exome platform matches the amino acid substitution and location in COSMIC

^bIn the gene column, italics indicates the variant found in the rat exome platform matches only the amino acid substitution in COSMIC

^cIn the gene column, bold, italics and underline indicates the variant found in the rat exome platform matches only the location in COSMIC

^dIn a cell line column, bold cell line indicates a homozygous mutation

^eIn a cell line column, italicized indicates a heterozygous mutation

^fIn a cell line column, Rat WES platform variant call/Sanger-based sequencing variant call. (WT: -, Variant: +). Bold and italics implies a sequencing discrepancy between the WES and Sanger

compare the mutation spectrum across each sample type [32]. Based on the mutation spectrum for each of the respective samples, we then focused on discriminating variants unique to each cell line and FF-FFPE samples and explored if there was a COSMIC signature relevant to a respective cell line indicating a high frequency of base pair mutations matching a specific tumor type. For the first comparison, we plotted dbSNP-registered variants along with all the exonic variants in the cell line and the FF-FFPE groups and observed similar frequency patterns across all samples with a very high frequency of C > T and T > C substitutions in the rat exome-seq data (Fig. 4a). Analysis of the mutation spectrum using the Kullback-Leibler method clustered dbSNP with the normal, FF-FFPE liver tissue and showed divergence of these groups from the tumor cell lines (Fig. 4b). We believe the bladder cell line, NBTII, clustered by itself due to the large number of respective variants identified with this cell line. Divergence plots focusing on the discriminating variants unique to each cell line, with the FF-FFPE variants filtered out, resulted in separation of the cell lines from dbSNP (Fig. 5a). Hierarchical clustering of the unique cell line specific exonic rat variants with the COSMIC signatures clustered the tumor cell line samples with the COSMIC signatures 16, 5, 8 and 3 (Fig. 5b). While COSMIC Signature 16 correlates to liver cancer and Signature 5 to all cancer types, both exhibit transcriptional strand bias for T > C mutations. Signature 8 associates with medulloblastomas, possibly linking an association with the chemically induced C6, glioblastoma cell line. Signature 3 aligned its pancreatic cancer phenotype with the chemically induced, pancreatic tumor cell line, DSL-6A/C1. The data lends support to a possible relationship between the chemically induced, tumor cell line variants and human COSMIC mutational signatures relative to a specific cancer phenotype.

Discussion

High throughput molecular-based methods like whole exome sequencing are needed to explore experimental in vitro systems and animal models for mechanisms and affected pathways in toxicity and disease. Yoshihara et al. recently published a probe design and application of a target capture sequencing of rat exons and conserved non-coding regions of 13 vertebrates [8]. Like the TargetEc probe design developed by Yoshihara, we employed merged RefSeq and Ensembl gene annotations to cover a 50 Mb exome region. However, in the final target capture design, Yoshihara added highly conserved, non-coding regulatory regions which tend to be shorter than exonic regions at about 5 bp in length. To compensate for the shorter regions (~5 bp), the probes were extended 100 bp to ensure efficient target region capture, which significantly increased the total target

capture size to 146.8 Mb. The larger probe library size may be viewed as a drawback since more reads would be required to obtain comparable sequencing depth for mutation detection in genomic regions of interest [8]. The current rat exome probe set at ~71 Mb in size focused on sequencing the protein coding regions plus UTRs. At half the size of the TargetEC platform, the probe set presents a more feasible approach to WES analysis in the rat.

The validity of exome sequencing relies on its ability to capture accurately targeted regions of interests. In evaluating the performance of a HTS platform, specific metrics looking at the target capture, alignment to the reference genome, variant calls and validation were taken into account. All performance metrics scored favorably with high capture sensitivity and specificity for the described rat exome platform design. This newly developed tool for rat exome sequencing makes feasible a comprehensive analysis of exonic variants across the entire rat exome and increases the possibilities for the discovery of gene mutations compared to candidate gene approaches. As expected, a diverse set of highly relevant, cancer-related genes were identified in the four rat tumor cell lines. Rat exome-sequencing targeted and captured gene mutations unique to each cell line and tumor type with a number of these genes previously reported in the literature for the respective human cancer type emphasizing the importance of rat cell lines in human translational research.

One of the first steps in analyzing sequencing data for probable candidate mutations is filtering the data for annotated, registered SNPs in dbSNP. The database reports far fewer SNPs for the rat compared to the mouse and human, emphasizing the point, that the rat genome is not currently as well annotated as its lab counterpart, the mouse. Our data set confirmed in all samples, the detection of annotated variants based on dbSNPs as well as the discovery of several non-annotated SNPs. Additional exome sequencing with this rat exome capture method can add valuable genomic data to rat databases.

Fresh frozen tissue is the “gold standard”, sample type of choice, for high throughput sequencing studies widely used to characterize variations from both normal and diseased tissues. Since these tissues do not undergo formalin fixation and processing at high temperatures, the DNA integrity is not altered. Based on capture sensitivity and specificity, the findings from this study show FFPE sample performance on the rat WES platform were comparable to FF tissue. Even though differences in the capture sensitivity and specificity existed between the two groups, the differences were marginal. At a coverage depth of 50X, our results affirmed FFPE tissues with a mean capture specificity of 57%. Capture specificity for FFPE samples evaluated at less stringent, but still strong

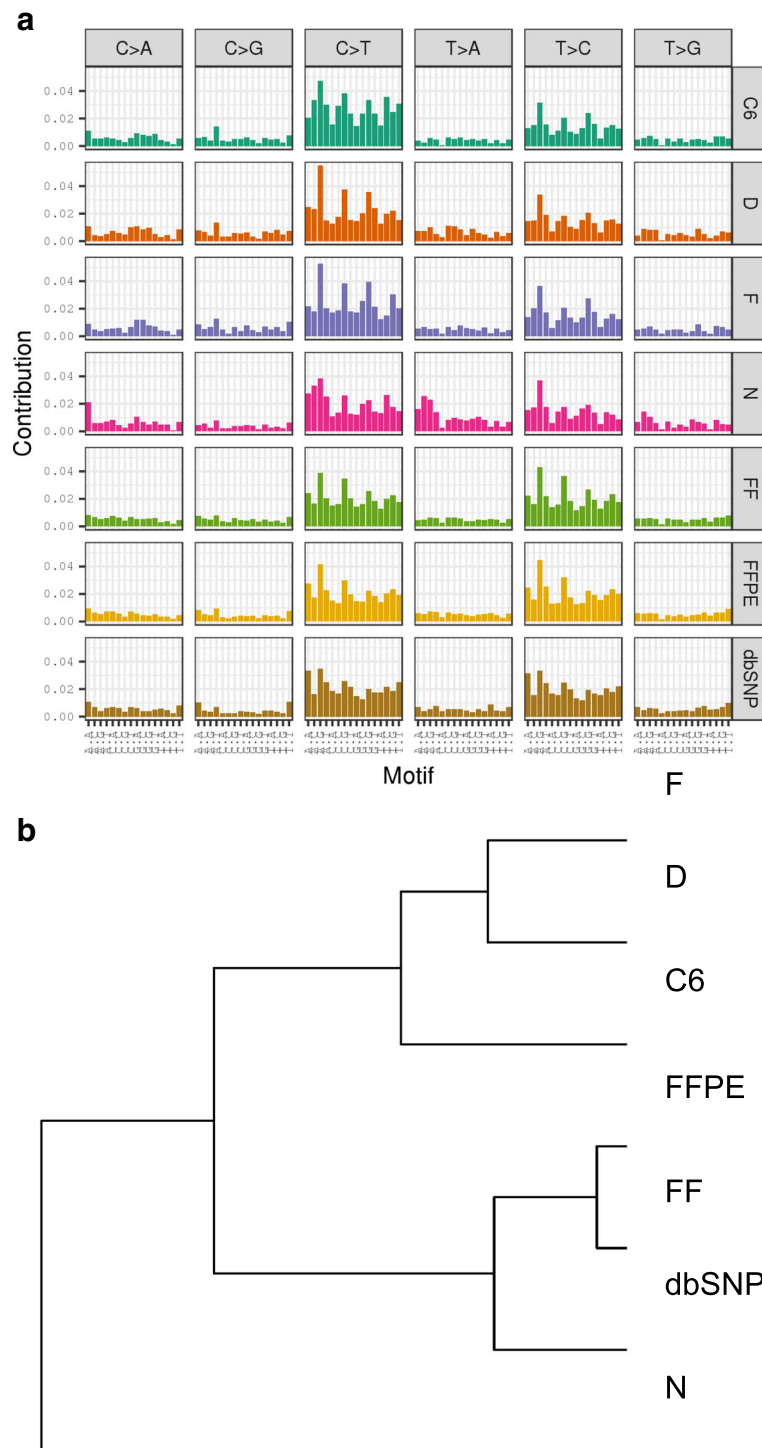
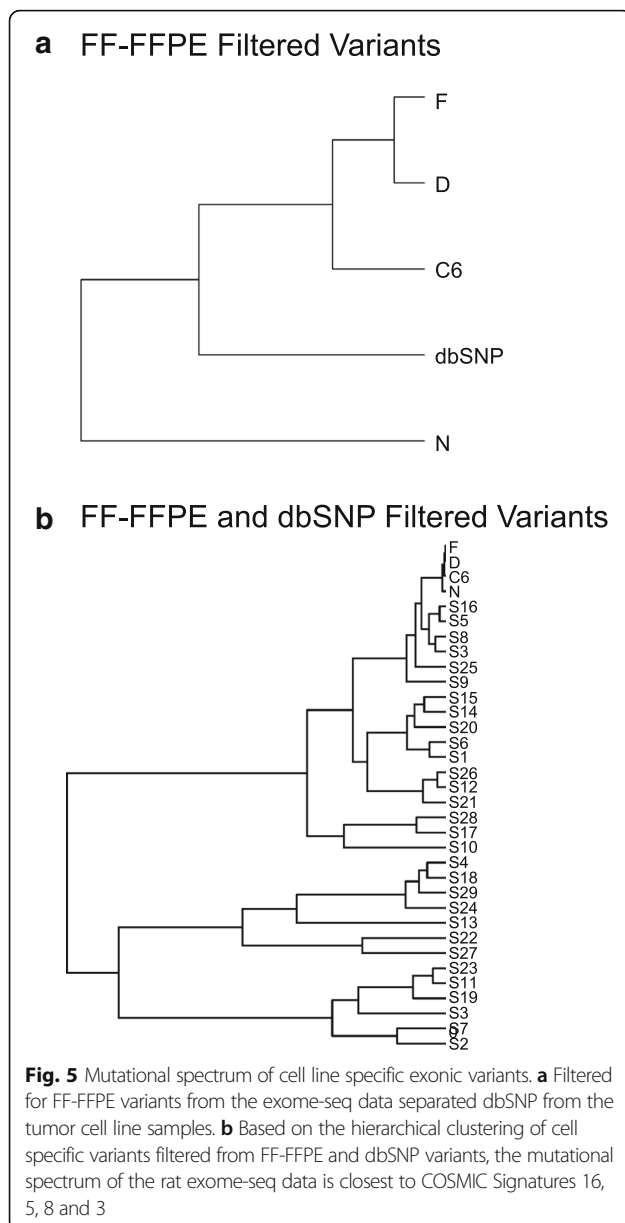


Fig. 4 Mutational spectrum of the rat exome-seq data. **a** All exonic variants captured across all samples in the 71 Mb design (50 Mb + UTRs) plus the dbSNP variants were plotted using the Kullback-Leibler divergence. **a** A high frequency of C > T and T > C mutations presented with minimal observed differences in the mutational spectrum across all samples and dbSNP. **b** Hierarchical clustering grouped dbSNP with the normal, FF-FFPE tissue and showed divergence of these groups from the tumor cell lines

coverage at an average sequencing depth of 30X resulted in an improved reference genome coverage to ~77%. As processed under conditions described herein, FFPE

samples were within the limits of achieving sequencing results comparable to FF tissue. Numerous clinical studies compared FF and FFPE specimens from normal



tissue and from various tumor types and achieved concordant, reliable sequencing results [33–35]. Although, caveats of nucleic acid degradation and protein cross-linking have been well described for FFPE tissue [34, 36], multiple sequencing studies have successfully demonstrated sufficient capture of target regions with acceptable capture sensitivity and specificity to obtain reliable sequencing data. We confirm the utility of using FFPE samples for exome-seq analysis in the current study.

COSMIC (Catalogue of Somatic Mutations in Cancer) is a public cancer database of human somatic mutations with data curated from the scientific literature and large-scale genomic studies from the Cancer Genome Project at the Sanger Institute [14]. Since the rat and its

cell lines are widely used in human translational disease studies, we matched variants identified in cancer-related genes found in our tumor cell line samples to reported mutations in these same genes in COSMIC. We selected a set of genes (*Tp53*, *Ncor1* and *Pik3ca*) common to multiple cancer types and evaluated the concordance of the captured exonic sequence by Sanger-based sequencing. The probe set we designed matched mutations with the identical amino acid substitutions found in the COSMIC database. In addition, qPCR analysis confirmed homozygous deletions of *Cdkn2a* and *Cdkn2b*, another common deletion present in multiple human cancers. The successful validation and matching of driver mutations found in common cancer genes suggests the rat WES probe set described here should be an immensely effective tool and impact human cancer genomic translational studies.

Sanger dideoxy terminator sequencing is the widely accepted “gold standard” for validating variants found using NGS even with its limitations as an orthogonal test [37]. We performed extensive confirmation of variants identified by WES to confirm the platform’s performance and ability to capture a wide-spectrum of mutations. Compared to other experimental animal models for which a probe set was designed to capture the exome, we assessed variant concordance between our rat exome method and Sanger sequencing on a sizeable set of variants to assure the accuracy of our in-silico probe design approach [8, 38, 39]. Our validation rate compared very favorably with Sanger data and represents a high degree of accuracy for the rat exome capture and sequencing platform.

When conducting experimental animal studies, accounting for strain differences within a species is an important aspect of the study design. In spite of the strain divergence among the cell line and tissue samples in our study, the total number of reads mapped well to the reference genome of Rn6, which is from the BN/SsNHsdMCW (Brown-Norway) strain. The percentage of aligned reads for the cell line and fresh frozen samples was 99%. The FFPE alignment was 90%. Although the probe sequence design was based on the Brown-Norway reference strain, and cell line variant comparisons were not matched to the respective rat strain, our designed probe set worked successfully for inbred and outbred laboratory rat strains for the identification of cancer-related genes, consistent with COSMIC annotation.

Human oncology studies employing genomic or exome sequencing methods often compare mutational profiles to defined cancer-associated mutational signatures in the COSMIC database. This allows insight into somatic mutations shared by a population and mechanisms driving cancer. Using tissue, kidney cell lines and embryonic stem cell

lines from *Fhit* knockout mice, Volinia et al. compared the mutation profile to signatures in the COSMIC database and established that a loss of gene expression controlled development of the ubiquitous signature 5 mutations in human cancer [40]. This discovery was based on the exome sequencing of the mouse, a highly annotated experimental animal model. For the current study, hierarchical clustering of four rat tumor cell lines coincided with defined molecular and cancer tissue signatures in the COSMIC database. Further work will continue to improve annotation of the rat genome and better establish the relationship between chemically induced rat tumors and defined human cancer-associated signatures in the COSMIC database.

Conclusions

A whole exome capture probe set and NGS sequencing of the rat exome represents a whole genome sequencing approach for investigators working in drug discovery, toxicology and hazard assessment. Targeted exome enrichment and sequencing can efficiently lead to variant discovery for insights into toxicity or disease etiologies, such as non-neoplastic and neoplastic lesions, along with developmental, reproductive, neurological, metabolic and endocrine disorders. The rat is an important experimental animal model used extensively in academia, industry and governmental agencies for toxicity and carcinogenicity assessment of various drugs, chemicals and hazardous agents. This newly developed rat exome enrichment system will expand the NGS tools available for rat genomic research.

Additional files

Additional file 1: Table S1. Rat exome-seq 500 flanking bp primer design for validation. (XLSX 20 kb)

Additional file 2: Table S2. PCR forward and reverse primer pairs for Sanger sequencing. (XLSX 10 kb)

Additional file 3: Figure S1. Exonic Variant Read Distribution. The distribution of all exonic variant reads across all chromosomes is directly proportional to the number of RefSeq genes for each chromosome, except chromosome 20. (PDF 15 kb)

Abbreviations

COSMIC: Catalogue of Somatic Mutations in Cancer; D: DSL-6 A/C1; dbSNP: Single Nucleotide Polymorphism database; F: FAT7; FF: Fresh frozen; FFPE: Formalin-fixed paraffin-embedded; HTS: High throughput screening; INDEL: Insertion/deletion; M: Million; N: NBTI; NGS: Next generation sequencing; SNV: Single nucleotide variant; UTR: Untranslated regions; WES: Whole exome sequencing; WT: Wild type

Acknowledgements

The authors wish to express their gratitude to Dr. Kevin Gerrish and Dr. Alison Harrill for review of the manuscript.

Funding

Funding for this research project was provided by the DNTP and DIR internal research funds. NIEHS Contract HHSN273201500005I provided funding for Sanger sequencing validation. NIEHS contract HHSN273201700001C provided bioinformatics support.

Availability of data and materials

The datasets used and/or analyzed during the current study are available in the GEO repository (<https://www.ncbi.nlm.nih.gov/sra/?term=SRP133156>).

Authors' contributions

JFF, DPP, OH, SH, VM, GGS, RRS, ARP, RAH, RCW, BAM made substantial contributions to the conception, design and acquisition of data for the study. DPP, RSS, OH, contributed to the probe set design. OH, SH, VM contributed to the library construction. GGS, JFF, BAM obtained services through the NIEHS Epigenomics Core Laboratory for WES. DPP, RSS, JFF and BAM contributed to the WES data analysis and interpretation. JF and SR completed the tissue and cell line experiments. AM, KH, KK, CL designed the selected gene targets and conducted the Sanger sequencing, qPCR analysis and data interpretation. ARP, RAH, RCS and JFF reviewed the liver histopathology. All the co-authors gave final approval for publication of the final version of the manuscript.

Ethics approval

Experiments were performed according to the guidelines established in the NIH Guide for the Care and Use of Laboratory Animals (National Research Council, 2011). Animals were treated humanely for alleviation of potential suffering, as approved by the National Institute of Environmental Health Sciences Animal Care and Use Committee.

Competing interests

JFF, SR, GGS, RAH, RCS and BAM are employees of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), of the United States Government. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of NIEHS, NIH or the United States Government. DPP and RRS are Bioinformaticians at Sciome, LLC and performed exome sequence location and annotation under NIEHS contract HHSN273201700001C under NIEHS supervision. OH, VM and SH comprised the team from Agilent Technologies that formulated the probe library for exome enrichment. JFF, SR, GGS, RAH, RCS and BAM supplied biological materials and isolated DNA to test the probe library for exome enrichment provided by Agilent. The NIEHS Epigenomics core facility (GGS) performed the exome sequencing. KH, KK, CL and AM are researchers at the Genomics Laboratory of LabCorp-Covance performed Sanger validation of targeted sequence variants under NIEHS contract HHSN273201500005I under NIEHS supervision. Neither Sciome, LabCorp-Covance or Agilent contributed financial funding for publication of this work. The authors adhere to policies on public sharing of data and materials.

Author details

¹Biomolecular Screening Branch, National Institute of Environmental Health Sciences, 111 T.W. Alexander Dr. Research Triangle Park, Durham, NC, USA. ²Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, Durham, NC, USA. ³Cellular and Molecular Pathology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, Durham, NC, USA. ⁴Sciome, LLC, Research Triangle Park, Durham, NC, USA. ⁵Agilent Technologies, Santa Clara, CA, USA. ⁶Covance Genomics Laboratory, Redmond, WA, USA. ⁷Adaptive Biotechnologies, Seattle, WA, USA.

Received: 26 February 2018 Accepted: 6 June 2018

Published online: 20 June 2018

References

- Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the brown Norway rat yields insights in to mammalian evolution. *Nat.* 2004;428:493–521.
- Sollner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stopka E, et al. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci Data.* 2017;4:170–85.
- Shimoyama M, Smith JR, Bryda E, Kuramoto T, Saba L, Dwinell M. Rat genome and model resources. *ILAR J.* 2017;58:42–58.
- Shimoyama M, Laulederkind SJ, De Pons J, Smith JR, Tutaj M, Petri V, Hayman GT, et al. Exploring human disease using the rat genome database. *Dis Model Mech.* 2016;9:1089–95.
- Meek S, Mashimo T, Burdon T. From engineering to editing the rat genome. *Mamm Genome.* 2017;28:302–14.

6. Tatreatult M, Bareke E, Nadaf J, Alirezaie N, Majewski J. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert Rev Mol Diagn.* 2015;15:749–60.
7. Wang Q, Shashikant CS, Jensen M, Altman NS, Girirajan S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep.* 2017;7:885–96.
8. Yoshihara M, Saito D, Sato T, Ohara O, Kuramoto T, Suyama M. Design and application of a target capture sequencing of exons and conserved non-coding sequences for the rat. *BMC Genomics.* 2016;17:593.
9. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* 2014;42:D764–70.
10. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics.* 2009;25:1754–60.
11. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
12. Cingolani P, Platts A, Le L W, 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;6:80–92.
13. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single Nucleotide variants. *Bioinformatics.* 2015;31:3673–5.
14. COSMIC, Catalogue of Somatic Mutations in Cancer <http://cancer.sanger.ac.uk/cosmic>. Accessed 22 Aug 2017.
15. Teer JK, Lori L, Bonnycastle PS, Chines NF, Hansen NA, Amy J, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* 2010;20:1420–31.
16. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. *Genome Biol.* 2011;12:R97.
17. Zhang P, Lehmann BD, Shyr Y, Guo Y. The utilization of formalin-fixed-paraffin-embedded specimens in high throughput genomic studies. *Int J Genomics.* 2017;19:26304.
18. Allen N. Biochemical study of tumors of the nervous system. In: Marks N, editor. *Research methods in neurochemistry.* New York: Plenum Press; 1978. p. 3–55.
19. Grobbsen B, De Deyn PP, Slegers H. Rat C6 glioma as experimental model system for the study of glioblastoma growth and invasion. *Cell Tissue Res.* 2002;10:257–70.
20. Barth RF, Kaur B. Rat brain tumor models in experimental neuro-oncology: the C6, 9L, T9, RG2, F98, BT4C, RT-2 and CNS-1 gliomas. *J Neuro-Oncol.* 2009;94:299–312.
21. Schlegel J, Piontek G, Kersting M, Schuermann M, Kappler R, Scherthan H, et al. The p16/Cdkn2a/Ink4a gene is frequently deleted in nitrosourea-induced rat glioma tumors. *Pathobiol.* 1999;67:202–6.
22. Asai A, Miyagi Y, Sugiyama A, Gamanuma M, Hong SI, Takamoto S, et al. Negative effects of wild-type p53 and s-Myc on cellular growth and tumorigenicity of glioma cells - implication of the tumor suppressor genes for gene therapy. *J Neuro-Oncol.* 1994;19:259–68.
23. NTP (National Toxicology Program). Report on carcinogens, fourteenth edition. Formaldehyde 2016;2000.
24. Reico L, Sisk S, Pluta L, Bermudez E, Gross EA, Chen Z, et al. P53 mutations in formaldehyde-induced nasal squamous cell carcinomas in rats. *Cancer Res.* 1992;52:6113–6.
25. Recio L. Oncogene and tumor suppressor gene alterations in nasal tumors. *Mutat Res.* 1997;380:27–31.
26. Bermudez E, Chen Z, Gross EA, Walker CL, Recio L, Pluta L, et al. Characterization of cell lines derived from formaldehyde-induced nasal tumors in rats. *Mol Carcinog.* 1994;9:193–9.
27. Liber HL, Benforado K, Crosby RM, Simpson D, Skopek TR. Formaldehyde-induced and spontaneous alterations in human hprt DNA sequence and mRNA expression. *Mutat Res.* 1989;226:31–7.
28. Kawanishi M, Matsuda T, Yagi T. Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation. *Front Environ Sci.* 2014;2:36.
29. Pettengill OS, Faris RA, Bell RH, Kuhlmann ET, Longnecker DS. Derivation of ductlike cell lines from a transplantable acinar cell carcinoma of the rat pancreas. *Am J Pathol.* 1993;143:292–303.
30. La Rosa S, Sessa F, Capella C. Acinar cell carcinoma of the pancreas: overview of Clinicopathologic features and insights into the molecular pathology. *Front Med.* 2015;2:1–13.
31. Jiao Y, Yonescu R, Offerhaus GJA, Klimstra DS, Maitra A, Eshleman JR, et al. Whole-exome sequencing of pancreatic neoplasms with acinar differentiation. *J Pathol.* 2014;232:428–35.
32. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat.* 1951;22:79–86.
33. Oh E, Choi YL, Kwon MJ, Kim RN, Kim YJ, Song JY, et al. Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples. *PLoS One.* 2015;10:e0144162.
34. Munchel S, Hoang Y, Zhao Y, Cottrell J, Klotzle B, Godwin AK, et al. Targeted or whole genome sequencing of formalin-fixed tissue samples: potential applications in cancer genomics. *Oncotarget.* 2015;6:25943–61.
35. De Paoli-Iseppi R, Johansson PA, Menzie AM, Dias KR, Pupo GM, Kakavand H, et al. Comparison of whole-exome sequencing of matched fresh and formalin-fixed paraffin-embedded melanoma tumours: implications for clinical decision making. *Pathol.* 2016;48:261–6.
36. Einaga N, Yoshida A, Noda H, Suemitsu M, Nakayama Y, Sakurada A, et al. Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation. *PLoS One.* 2017;12:e0176280.
37. Beck TF, Mullikin JC, NISC Comparative Sequencing Program, Biesecker LG. Systematic evaluation of sanger validation of next-generation sequencing variants. *Clin Chem.* 2016;64:647–54.
38. Robert C, Fuentes-Utrilla P, Troup K, Loecherbach J, Turner F, Talbot R, et al. Design and development of exome capture sequencing for the domestic pig (*Sus scrofa*). *BMC Genomics.* 2014;15:550.
39. Broeckx BJ, Coopman F, Verhoeven GE, Bavegems V, De Keulenaer S, De Meester E, et al. Development and performance of a targeted whole exome sequencing enrichment kit for the dog (*Canis Familiaris* build 3.1). *Sci Rep.* 2014;4:5597.
40. Volinia S, Druck T, Paisie CA, Schrock MS, Huebner K. The ubiquitous 'cancer mutation signature' 5 occurs specifically in cancers with deleted *FHIT* alleles. *Oncotarget.* 2017;8:102199–211.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

