Topic Introduction

# Cross-Species Analysis of Mouse and Human Cancer Genomes

Carla Daniela Robles-Espinoza and David J. Adams[1]

*Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1HH, United Kingdom*

Fundamental advances in our understanding of the human cancer genome have been made over the last five years, driven largely by the development of next-generation sequencing (NGS) technologies. Here we will discuss the tools and technologies that have been used to profile human tumors, how they may be applied to the analysis of the mouse cancer genome, and the results thus far. In addition to mutations that disrupt cancer genes, NGS is also being applied to the analysis of the transcriptome of cancers, and, through the use of techniques such as ChIP-Seq, the protein–DNA landscape is also being revealed. Gaining a comprehensive picture of the mouse cancer genome, at the DNA level and through the analysis of the transcriptome and regulatory landscape, will allow us to "biofilter" for driver genes in more complex human cancers and represents a critical test to determine which mouse cancer models are faithful genetic surrogates of the human disease.



## INTRODUCTION

Mice have been used as models of human cancer for well over a century since Abbie Lathrop and Leo Loeb first described similarities between skin tumors forming in mice and those forming in humans (reviewed in Rader 2004). The essence of these comparisons, and most comparisons made between mouse and human tumors to this day, is that a good tumor model is one that produces tumors that a histopathologist finds indistinguishable from the cognate human tumor type. In recent years, however, extensive analysis of human tumors as part of efforts such as the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium et al. 2010) and The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network 2008; see Table 1) has revealed significant genetic intratumor heterogeneity. Indeed, tumors of the same histology may not carry any mutated genes in common (Stephens et al. 2012), suggesting that there are many paths toward malignancy.

Mouse models of cancer are being used to understand how tumors and the immune system interact as preclinical models of human cancer and for other translational activities. For these models to be faithful, and to aid in their interpretation, it is important that in addition to an understanding of each model's pathology we also understand the underlying genetic mutations that drive tumor evolution. Because most tumor models are generated via the conditional loss of tumor-suppressor genes or activation of oncogenes, the initiating mutations and their timing are often known. This has major advantages in terms of the reproducibility of tumor formation and presentation, but it

[1]Correspondence: da1@sanger.ac.uk


Advance Online Article. Cite this introduction as *Cold Spring Harb Protoc*; doi:10.1101/pdb.top078824

**TABLE 1.** An overview of the current International Cancer Genome (ICGC) and The Cancer Genome Atlas (TCGA) projects

| | ICGC | | |
|---|---|---|---|
| Tumor type | Number of donors | Number of samples | Lead jurisdiction (funding organizations) |
| **Bladder** | | | |
| Invasive urothelial bladder cancer | Data available for 68 | Data available for 344 | USA (NCI/NHGRI/NIH) |
| **Bone** | | | |
| Osteosarcoma | | Proposed: 250 | United Kingdom (WT/SCAT/BCRT/EuroBoNeT) |
| Chondrosarcoma | | Proposed: 200 | |
| Rare subtypes | | Proposed: 50 | |
| **Breast** | | | |
| ER-positive, *HER2*-negative | | Proposed: 500 | European Union/United Kingdom (European Commission FP7) |
| Amplification of the *HER2* gene | | – | France |
| Triple negative/lobular/other | Data available for 117 | Proposed: 500; data available for 117 | United Kingdom (WT/BBC) |
| Ductal carcinoma | | Proposed: 138 | Mexico (CSHI) |
| Ductal and lobular | Data available for 865 | Data available for 4603 | USA (NCI/NHGRI/NIH) |
| Asian phenotype | | – | South Korea |
| **Central nervous system** | | | |
| Medulloblastoma and pediatric pilocytic astrocytoma | Data available for 135 | Data available for 121 | Germany (BMBF/DKH) |
| Glioblastoma multiforme | Data available for 565 | Data available for 5606 | USA (NCI/NHGRI/NIH) |
| Pediatric medulloblastoma | Data available for 19 | Proposed: Genomic 300, RNA 1000; data available for 19 | Canada (GC/GBC/TFRI/SAL/CRF/GFCC/OICR/POGO/CGP/CHSBTC/MCHF/BRAINC) |
| Lower grade glioma | Data available for 161 | Data available for 376 | USA (NCI/NHGRI/NIH) |
| **Cervix** | | | |
| Cervical squamous cell carcinoma | Data available for 31 | Data available for 55 | USA (NCI/NHGRI/NIH) |
| **Colorectal** | | | |
| Colon-adenocarcinoma | Data available for 423 | Data available for 2056 | USA (NCI/NHGRI/NIH) |
| Adenocarcinoma, non-Western | | – | China (CCGC/MST/HTRDP/NNSFC) |
| Rectal-adenocarcinoma | Data available for 168 | Data available for 740 | USA (NCI/NHGRI/NIH) |
| **Esophagus** | | | |
| Esophageal adenocarcinoma | | – | United Kingdom (CRUK) |
| Squamous carcinoma | | – | China (CCGC/MST/HTRDP/NNSFC) |
| **Head, neck, and nasopharynx** | | | |
| Gingivobuccal | | – | India (DBMST) |
| Squamous cell carcinoma | Data available for 263 | Data available for 1568 | USA (NCI/NHGRI/NIH) |
| Squamous cell carcinoma of oral cavity/oropharynx/sinonasal cavity/hypopharynx/larynx | Proposed: 92 | | Mexico (CSHI); partner USA (NCI/STARRCC) |
| Thyroid carcinoma | Data available for 193 | Data available for 648 | USA (NCI/NHGRI/NIH) |
| Papillary thyroid carcinoma | | – | Saudi Arabia (KFSHRC) |
| Nasopharyngeal carcinoma, Asia | | – | China (CCGC/MST/HTRDP/NNSFC) |
| **Hematopoietic and lymphoid tissues** | | | |
| Chronic lymphocytic leukemia with mutated and unmutated IgVH | Data available for 177 | Proposed: 500; data available for 274 | Spain (IHC/SMSI) |
| Acute myeloid leukemia | Data available for 200 | Data available for 561 | USA (NCI/NHGRI/NIH) |
| Germinal center B-cell-derived lymphomas | Data available for 10 | Data available for 10 | Germany (BMBF) |
| Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies | Data available for 129 | Data available for 129 | United Kingdom (WT/KKLF) |
| Diffuse large B-cell lymphoma | Proposed: 100 | | Mexico (CSHI); partner USA (NCI) |
| Acute myeloid leukaemia | | – | South Korea (NPPGM) |
| **Kidney** | | | |
| Renal cell carcinoma (focus on but not limited to clear cell subtype) | | – | European Union/France (European Commission FP7) |

(*continued*)

**TABLE 1.** *Continued*

| | ICGC | | |
|---|---|---|---|
| Tumor type | Number of donors | Number of samples | Lead jurisdiction (funding organizations) |
| Clear cell carcinoma | Data available for 502 | Data available for 3167 | USA (NCI/NHGRI/NIH) |
| Papillary carcinoma | Data available for 95 | Data available for 357 | USA (NCI/NHGRI/NIH) |
| **Liver** | | | |
| Hepatocellular carcinoma (secondary to alcohol and adiposity) | Data available for 125 | Proposed: 514; data available for 82 | France (INCa) |
| Hepatocellular carcinoma (virus associated) | Data available for 265 | Data available for 267 | Japan (NIBI/RIKEN) |
| Hepatocellular carcinoma | Data available for 62 | Data available for 197 | USA (NCI/NHGRI/NIH) |
| Hepatocellular carcinoma (HBV associated) | | – | China (CCGC/MST/HTRDP/NNSFC) |
| **Lung** | | | |
| Adenocarcinoma | Data available for 292 | Data available for 1649 | USA (NCI/NHGRI/NIH) |
| Squamous cell carcinoma | Data available for 279 | Data available for 2475 | USA (NCI/NHGRI/NIH) |
| Small-cell lung carcinoma | Data available for 1 | Data available for 1 | United Kingdom |
| **Ovary** | | | |
| Serous cystadenocarcinoma | | – | Australia (NHMRC/QSG/UQ/IMB) |
| Serous cystadenocarcinoma | Data available for 576 | Data available for 8200 | USA (NCI/NHGRI/NIH) |
| **Pancreas** | | | |
| Ductal adenocarcinoma | Proposed: 375 | – | Australia (NHMRC/QSG/UQ/IMB/CCNSW/GIMR) |
| Ductal adenocarcinoma | Data available for 75 | Proposed: 500; data available for 95 | Canada (OICR/OMRI/CFI) |
| Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors | | Proposed: 250 | Italy (IMEUR/UV) |
| Adenocarcinoma | | – | USA (NCI/NHGRI/NIH) |
| **Skin** | | | |
| Cutaneous melanoma | Data available for 129 | Data available for 243 | USA (NCI/NHGRI/NIH) |
| Malignant melanoma | Data available for 1 | Data available for 1 | United Kingdom |
| **Stomach** | | | |
| Intestinal and diffuse type | Data available for 10 | Data available for 10 | China (CCGC/MST/HTRDP/NNSFC) |
| Adenocarcinoma | Data available for 159 | Data available for 621 | USA (NCI/NHGRI/NIH) |
| **Testis and prostate** | | | |
| Early onset | Data available for 9 | Proposed 250; data available for 9 | Germany (BMBF) |
| Adenocarcinoma | Data available for 127 | Data available for 602 | USA (NCI/NHGRI/NIH) |
| Adenocarcinoma | Data available for 10 | Proposed 500; data available for 28 | Canada (OICR/PCC) |
| Adenocarcinoma | Data available for 2 | Proposed 250; data available for 2 | United Kingdom (CRUK) |
| **Uterus** | | | |
| Uterine corpus endometrial carcinoma | Data available for 451 | Data available for 2200 | USA (NCI/NHGRI/NIH) |

Data obtained from the ICGC Data Portal (Zhang et al. 2011) release 10. Released data are available at ftp://data.dcc.icgc.org/.

The projects are organized by tumor type. The "number of donors" column indicates the number of patients from whom tumors have been obtained, and "number of samples" refers to the total available number of samples. "Proposed" indicates that the data are not yet available, but will be released in the future. The shaded projects are funded by the TCGA.

NCI, National Cancer Institute; NHGRI, National Human Genome Research Institute; NIH, National Institutes of Health; WT, The Wellcome Trust; SCAT, Skeletal Cancer Action Trust; BCRT, Bone Cancer Research Trust; BBC, Breakthrough Breast Cancer; CSHI, Carlos Slim Health Institute; BMBF, Federal Ministry of Education and Research; DKH, German Cancer Aid; GC, Genome Canada; GBC, Genome British Columbia; TFRI, Terry Fox Research Institute; SAL, Hospital for Sick Children, Sonia and Arthur Labatt Brain Tumor Research Centre; CRF, Hospital for Sick Children, Chief of Research Fund; GFCC, Hospital for Sick Children, Garron Family Cancer Centre; OICR, Ontario Institute for Cancer Research; POGO, Pediatric Oncology Group Ontario; CGP, Hospital for Sick Children, Cancer Genetics Program; CHSBTC, Clark H. Smith Brain Tumor Centre; MCHF, Montreal Children's Hospital Foundation; BRAINC, Hospital for Sick Children, B.R.A.I.N. Child; CCGC, Chinese Cancer Genome Consortium; MST, Ministry of Science and Technology; HTRDP, National High Technology Research and Development Program of China; NNSFC, National Natural Science Foundation of China; CRUK, Cancer Research UK; DBMST, Department of Biotechnology, Ministry of Science and Technology; STARRCC, STARR Cancer Consortium; KFSHRC, King Faisal Specialist Hospital and Research Centre; IHC, Institute of Health Carlos III; SMSI, Spanish Ministry of Science and Innovation; KKLF, Kay Kendall Leukaemia Fund; NPPGM, The National Project for Personalized Genomic Medicine; INCa, Institut National du Cancer; NIBI, National Institute of Biomedical Innovation; NHMRC, National Health and Medical Research Council; QSG, Queensland State Government; UQ, University of Queensland; IMB, Institute for Molecular Bioscience; CCNSW, Cancer Council New South Wales; GIMR, Garvan Institute of Medical Research; OMRI, Ontario Ministry of Research and Innovation; CFI, Canada Foundation for Innovation; IMEUR, Italian Ministry of Education, University and Research; UV, University of Verona; PCC, Prostate Cancer Canada.

also represents a somewhat unnatural beginning, and thus it is unclear whether mouse tumors follow the same paths toward malignancy as human tumors. In an analogous manner, it is also unclear whether mouse tumors acquire the same constellations of mutations as human cancers. Furthermore, it is well known that mouse cells rarely undergo telomere crisis (Maser et al. 2007), and that there are species differences between mouse and human cells in terms of their requirements for transformation (Rangarajan et al. 2004). Mouse chromosomes are also acrocentric and hence structurally different to human chromosomes (Mouse Genome Sequencing Consortium et al. 2002). Collectively, these differences may have important ramifications for the genes mutated in mouse tumor models with the mutation spectra also influencing tumor growth, the likelihood that a tumor will metastasize, its response to therapy, and other important aspects of tumor biology. Here we will discuss the tools and technologies being applied to analyze cancer genomes and outline the progress made to date in profiling mouse cancers using these tools.

## AN OVERVIEW OF SEQUENCING TECHNOLOGIES FOR CANCER GENOME ANALYSIS

Since the completion of the Human Genome Project in April 2003 (International Human Genome Sequencing Consortium 2004), DNA sequencing technologies have evolved at an unprecedented pace. At the time of this publication, the costs of sequencing a human genome had dropped by more than 10,000-fold in a little over a decade (Wetterstrand), with emerging developments promising to bring us closer to the era of personalized medicine.



### DNA Sequencing

With the transition from classical chain-termination methods, which were initially described by Frederick Sanger (Sanger et al. 1977) and used to assemble the draft of the human genome, to NGS technologies, it became possible to sequence thousands of DNA samples in parallel. With these methods genomic DNA is broken into small pieces, amplified, and then sequenced before being aligned to a reference genome, or assembled de novo. NGS technologies operate on the principle of sequencing by synthesis, which relies on the real-time detection of the order in which specific nucleotides are incorporated by a polymerase while a DNA strand is replicated.

Once fragments have been sequenced, several analyses to obtain meaningful information need to be performed. First, the DNA fragments that have been sequenced need to be mapped to an existing reference genome. Software such as the Burrows–Wheeler Aligner may be used (Li and Durbin 2009). Having aligned all reads from DNA fragments to a reference genome, differences between tumor and the germline sample pairs need to be identified. Software such as SAMtools (Li 2011) and the Genome Analysis Toolkit (DePristo et al. 2011) suite, as well as other commonly used algorithms like SomaticSniper (Larson et al. 2012) and VarScan 2 (Koboldt et al. 2012) may be used to identify somatically altered nucleotide positions. Structural variants, be they deletions, insertions, or mobile genetic element insertions, may be identified using discordantly mapped reads (Campbell et al. 2008). BreakDancer (Chen et al. 2009), SVMerge (Wong et al. 2010), and HYDRA (Quinlan et al. 2010) are software tools designed to find structural variants. Tumor cellularity and ploidy need to be considered when calling somatic, variants, and rare polymorphic variants should be removed.

### RNA Sequencing

Apart from identifying DNA variants present in tumor genomes, analysis of the transcriptome, which captures information on which genes are over- or underexpressed, or differentially processed, is often informative. This approach involves sequencing pools of complementary DNA (cDNA) fragments and mapping them back to the reference genome in a slice-aware manner. Several algorithms including TopHat (Trapnell et al. 2009) and STAR (Dobin et al. 2013) may be applied for this purpose. The number of reads mapping to a gene is then used to infer its transcriptional activity, whereas read-pair mapping information is used to look for differences in RNA processing such as splicing, and cancer

fusion genes. The reader is referred to Wang et al. (2009) for a description of the RNA sequencing approach, its challenges, and its future prospects.

### Sequencing to Find DNA–Protein Interactions

Up until the advent of NGS technologies, one of the most widely used methods to investigate the global landscape of DNA–protein interactions was chromatin immunoprecipitation coupled with microarray chips, known as ChIP-chip (Bulyk 2006). More recently, ChIP-Seq has been developed allowing DNA–protein interactions to be profiled with near base-pair resolution (Kulakovskiy et al. 2013). For a detailed explanation of the methodology and a discussion of its challenges and advantages, see Park (2009).

### Single-Cell DNA and RNA Sequencing and Other Emerging Technologies

Even though the study of cancer genomes has been enormously facilitated by NGS technologies, until recently these approaches used DNA derived from bulk cells or tissue; as a consequence, information on intratumor heterogeneity and clonality is essentially lost. Recently, reliable single-cell whole-genome, exome-, and transcriptome-sequencing methods have been developed (Lu et al. 2012; Zong et al. 2012). These methods can be used to profile the somatic changes, copy number variants, and gene expression in single tumor cells to build a picture of the clonal architecture of a cancer.



## THE CANCER GENOME OF HUMANS AND MICE

Analysis of human cancer genomes using the technology and software tools described above has revealed considerable complexity both in terms of the nucleotide landscape and in terms of chromosome copy number and rearrangements, with aneuploidy being pervasive in most tumor types. Table 1 outlines the current ICGC and TCGA projects that are being performed or have been completed using NGS technology. The same tools used for these projects are now being applied to mouse cancer models. The rationale for these experiments is to provide a shortcut to the identification of genes relevant for disease pathogenesis in human cancer and also to validate that mouse models of cancer faithfully recapitulate the human disease. Unlike the sequencing of complex human cancer genomes, the sequencing of tumors from mice, which are derived from inbred lines that are homozygous at every locus or that are derived from a common pool of founders, should be a much simpler task. Furthermore, mouse models of cancer are generally engineered to contain potent driver mutations and tumors in these models form with a short latency; thus the expectation is that the ratio of driver to passenger mutations will be higher (Frese and Tuveson 2007). Because experiments with mouse models of cancer can also be easily scaled, sample size is rarely a limiting factor and large collections of tumors may be analyzed even for rare or hard-to-obtain tumor types.

Pilot mouse cancer sequencing experiments have thus far produced promising results. For example, paired-end sequencing of mammary tumors from mice carrying conditional alleles of *Trp53* alone or in combination with *Brca1* or *Brca2* alleles (modeling hereditary breast cancer) or conditional *Cdh1* (a model of lobular breast cancer) showed key features associated with the cognate human tumor type (Varela et al. 2010). Namely, a homozygous in-frame deletion of *Lrp1b* was found, a feature shown to be present in ~5% of human tumor-cell lines. Inter- and intrachromosomal rearrangements were also observed but were fewer in number than those found in human breast cancers, which is likely to be caused by their relatively short evolution time (i.e., they have not had the opportunity to acquire a high mutational load). Importantly, significant heterogeneity in the patterns of somatic rearrangement was also observed. Because heterogeneity contributes to disease progression and response to therapy and represents an important aspect of human mammary tumorigenesis, it is important to highlight the fact that as with human tumors, the genomes of mouse tumors do not all follow a common path.

As further evidence of the promise of this approach, sequencing the genome of a tumor:normal pair from a transgenic mouse expressing a *PML-RARα* fusion oncogene modeling acute promyelocytic leukemia revealed a somatic *Jak1*$^{V657F}$ mutation in the same residue of *JAK1* as previously found in human acute promyelocytic/lymphoblastic leukemias (Wartman et al. 2011). Furthermore, somatic deletion of *Utx* (a histone H3 lysine 27 demethylase) was observed, and deletion of this gene was subsequently found to occur in human *PML-RARα* cases. These initial studies suggest that sequencing mouse tumor genomes is a relevant approach for the discovery of driver mutations found in human malignancies. Some differences, however, have become apparent. For example, sequencing of mouse mammary tumors did not reveal the tandem duplication events that had previously been observed in human breast cancers (Stephens et al. 2012). Subtleties such as this illustrate that there may be important differences between the mouse and human genome and these differences need to be understood.

## SEQUENCE TO DRIVER MUTATIONS, DATA INTEGRATION, AND BIOFILTERING

One key aspect of sequencing genomes from a model organism and then comparing the results to the human genome is the requirement to "lift" data between assemblies. Although the majority of mouse genes are conserved in human, with some estimates up to 95%, there are some genes that are species-specific (Guigo et al. 2003). Some gene families have also been expanded or contracted through evolution (Demuth et al. 2006; Krushna Padhi et al. 2006). A particularly notable example is the expansion of olfactory receptors in mouse (Wynn et al. 2012) when compared to human (Rouquier et al. 2000). Other genes such as the Ubiquitin-specific protease 6 (*USP6*) are apparently absent from the mouse genome and found in primates (Flicek et al. 2013). Some of these differences may be due to gene annotation differences, whereas others are the result of divergence. Thus mapping or lifting data from the mouse to human genome, or comparing mouse and human genes, can be challenging.

Several tools have been developed to facilitate the task of comparing data between genomes. One tool is called Compara and is part of the suite of tools provided by the Ensembl Genome Browser (www.ensembl.org) (Flicek et al. 2013). Compara allows comparisons between genomes, using DNA sequences, protein sequences, and synteny maps. It also provides information on gene orthologs and paralogs. Importantly, Ensembl provides an application programming interface that allows bioinformaticians to write scripts or software to interrogate the Compara database simplifying the analysis of large data sets. In addition to Compara, other tools such as Inparanoid (O'Brien et al. 2005) have been developed. The Inparanoid project gathers proteomes of completely sequenced eukaryotic species plus *Escherichia coli* and calculates pairwise ortholog relationships among them to identify ortholog pairs. Thus mutations identified by sequencing a mouse genome can be mapped to residues on a human protein and back again.

Another application for transferring mouse data to the human genome for cross-species comparisons is the UCSC LiftOver tool (Hinrichs et al. 2006). LiftOver converts genome coordinates and genome annotation files between assemblies. The current version supports both forward and reverse conversions, as well as conversions between selected species including between human and mouse. In addition to these tools for comparing genomes, several tools exist to predict the likely consequence of a particular mutation found in a cancer genome. When the variant identified is a deletion, or truncating mutation, the interpretation is relatively straightforward. However, changes such as nonsynonymous single-nucleotide variants can be harder to understand and interpret. One tool in common use is called PANTHER (Brunham et al. 2005), which uses evolutionary relationships to determine the likelihood that a nucleotide mutation will be deleterious. Other software tools include SIFT (Ng and Henikoff 2001) and PolyPhen-2 (Adzhubei et al. 2010), which use a combination of structural and comparative evolutionary considerations to score the likely effect of a mutation.

More recently, other approaches have been developed including InVEx, which uses a permutation-based method to calculate if there is an enrichment of mutations in coding sequences of a gene compared to flanking intronic DNA sequence (Hodis et al. 2012). In this way genes undergoing

selection can be identified. This application has the advantage of being able to account for local biases in mutation rates and may be particularly useful in carcinogen-driven mouse models where there are likely to be a large number of passenger mutations within coding sequences that may complicate the task of sorting drivers from passengers.

In addition to the mechanics of comparing data between species, it is important to bear in mind that genes dysregulated in a model system may be altered through a different biological mechanism than those observed in human tumors. For example, genes deleted in human tumors may be point-mutated in the mouse or silenced by changes in the epigenome. Likewise, genes that are recurrently amplified in human may be translocated in mouse. Thus those "biofiltering" between genomes to identify candidate cancer genes need to consider this complexity. Ultimately, in-laboratory validation of all genes identified as candidate drivers is required to confirm their contribution to tumorigenesis.

## EXPERIMENTAL DESIGN: ADVICE AND GUIDANCE!

Although the major challenge of most genome-sequencing endeavors is usually financial, the saying "Rubbish in, rubbish out" is also true. The ICGC and TCGA efforts to sequence human tumors have set very strict guidelines for sample quality for genome-sequencing experiments (International Cancer Genome Consortium 2010) and these guidelines are generally applicable to the analysis of mouse tumors. The ICGC/TCGA guidelines stipulate that tumors for DNA sequencing should be reviewed by an expert pathologist or team of pathologists; that each tumor be graded for tumor-cell content, the amount of stroma, and immune cell infiltrate; and should also be classified histopathologically. Ideally, a tumor should be at least 80% tumor cells, and selecting tumors that are not necrotic is also favored. DNA from fresh frozen tissue is also considered optimal.

The ICGC/TCGA guidelines also suggest that it is mandatory for matched germline DNA for each tumor to be available. Although the genome sequence of many commonly used laboratory mouse strains is now available (Keane et al. 2011; Wong et al. 2012), many investigators have colonies that have been bred in isolation for many generations and thus will have drifted from the founding stock. As rare germline variants that pass variant filters represent a major source of erroneous somatic calls, it is important that germline genomic DNA from each mouse is available. This is particularly the case when tumors from models containing several alleles are being analyzed, as many of these models are from backgrounds for which little breeding history is available. Tail DNA is an ideal source of control germline DNA, except where the tail is heavily infiltrated with leukemic or other tumor cells. If funds for sequencing are limiting, one viable strategy is to sequence the parents of the mice that were used to generate the cohort of animals placed on tumor watch, because most germline variants should be captured within the genomes of these mice and can be removed from the somatic call set. Our advice is to sequence only tumors where full knowledge of the life history of the tumor is known—for example, information on the tumor latency, the set of predisposing alleles, the breeding history of the mouse, and a full histology report.

Generally, at least 5 µg of DNA is required to sequence a whole exome or genome, depending on the method used. It is important to have at least that amount of DNA again so that downstream

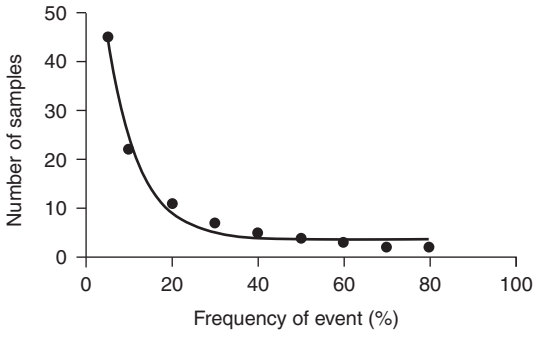


**FIGURE 1.** A power calculation for the number of sequenced tumors required to detect recurrently mutated cancer genes with 90% power. The assumptions made are that there is approximately one mutation per megabase, that the average coding length of a cancer gene is 2.5 kb, and that cancer genes are found with a prevalence of ≥5%. If these assumptions are met, an analysis of 50 cancers will yield the following true- and false-positive rates: TPR%, 90; FPR%, 5.

genotyping may be performed to validate any somatic mutations identified. For most transcriptome sequencing protocols, ~5 µg of whole RNA is required, although some protocols require less (Tariq et al. 2011; Sultan et al. 2012). Where possible, a useful source of material for follow-up studies may be cell lines derived from the tumor.

A common question that is asked at the initiation of a mouse cancer sequencing experiment is "How many tumors should be sequenced?" The answer to this question depends very much on the mutational spectra and tumor heterogeneity. As a general rule, around 50 tumors will capture most of the recurrently mutated genes within a mouse tumor model. Figure 1 provides a power calculation for the likelihood that mutations in a driver cancer gene will be identified when different numbers of tumors are analyzed.

## THE MOUSE CANCER GENOME AND BUILDING A BETTER MOUSE

Over the coming years, as sequencing technologies become cheaper and analysis tools more user-friendly, the genome sequences of hundreds if not thousands of mouse cancers will be produced. From this analysis, a clearer picture of the landscape of the mouse cancer genome will be revealed, and importantly a greater understanding of how well our models recapitulate the human disease will be gleaned. This analysis may tell us some home truths from which we will evolve and build better systems for studying the genes and mechanisms that that drive cancer development in mouse and man.



## REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249.

Brunham LR, Singaraja RR, Pape TD, Kejariwal A, Thomas PD, Hayden MR. 2005. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the *ABCA1* gene. *PLoS Genetics* **1:** e83.

Bulyk ML. 2006. DNA microarray technologies for measuring protein–DNA interactions. *Curr Opin Biotechnol* **17:** 422–430.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genetics* **40:** 722–729.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006. The evolution of mammalian gene families. *PLoS ONE* **1:** e85.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genetics* **43:** 491–498.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Frese KK, Tuveson DA. 2007. Maximizing mouse cancer models. *Nat Rev Cancer* **7:** 645–658.

Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci* **100:** 1140–1145.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res* **34:** D590–D598.

Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, et al. 2012. A landscape of driver mutations in melanoma. *Cell* **150:** 251–263.

International Cancer Genome Consortium. 2010. Section E.6—Quality standards of samples. In *Updates to goals, structure, policies & guidelines.* http://icgc.org/icgc/goals-structure-policies-guidelines.

International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, et al. 2010. International network of cancer genome projects. *Nature* **464:** 993–998.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22:** 568–576.

Krushna Padhi B, Akimenko MA, Ekker M. 2006. Independent expansion of the keratin gene family in teleostean fish and mammals: An insight from phylogenetic analysis and radiation hybrid mapping of keratin genes in zebrafish. *Gene* **368:** 37–45.

Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V. 2013. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J Bioinform Comput Biol* **11:** 1340004.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28:** 311–317.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, Zhu P, Hu X, Xu L, Yan L, et al. 2012. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338:** 1627–1630.

Maser RS, Choudhury B, Campbell PJ, Feng B, Wong KK, Protopopov A, O'Neil J, Gutierrez A, Ivanova E, Perna I, et al. 2007. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447:** 966–971.

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* **11:** 863–874.

O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33:** D476–D480.

Park PJ. 2009. ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genetics* **10:** 669–680.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20:** 623–635.

Rader K. 2004. *Making mice: Standardizing animals for American biomedical research, 1900–1955*. Princeton University Press, Princeton, NJ.

Rangarajan A, Hong SJ, Gifford A, Weinberg RA. 2004. Species- and cell type-specific requirements for cellular transformation. *Cancer Cell* **6:** 171–183.

Rouquier S, Blancher A, Giorgi D. 2000. The olfactory receptor gene repertoire in primates and mouse: Evidence for reduction of the functional fraction in primates. *Proc Natl Acad Sci* **97:** 2870–2874.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* **74:** 5463–5467.

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486:** 400–404.

Sultan M, Dokel S, Amstislavskiy V, Wuttig D, Sultmann H, Lehrach H, Yaspo ML. 2012. A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochem Biophys Res Commun* **422:** 643–646.

Tariq MA, Kim HJ, Jejelowo O, Pourmand N. 2011. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res* **39:** e120.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Varela I, Klijn C, Stephens PJ, Mudie LJ, Stebbings L, Galappaththige D, van der Gulden H, Schut E, Klarenbeek S, Campbell PJ, et al. 2010. Somatic structural rearrangements in genetically engineered mouse mammary tumors. *Genome Biology* **11:** R100.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genetics* **10:** 57–63.

Wartman LD, Larson DE, Xiang Z, Ding L, Chen K, Lin L, Cahan P, Klco JM, Welch JS, Li C, et al. 2011. Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression. *J Clin Invest* **121:** 1445–1455.

Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at http://www.genome.gov/sequencingcosts. Accessed 06/03/2013.

Wong K, Keane TM, Stalker J, Adams DJ. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol* **11:** R128.

Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DJ, Keane TM. 2012. Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biol* **13:** R72.

Wynn EH, Sanchez-Andrade G, Carss KJ, Logan DW. 2012. Genomic variation in the vomeronasal receptor gene repertoires of inbred mice. *BMC Genomics* **13:** 415.

Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, et al. 2011. International Cancer Genome Consortium Data Portal—A one-stop shop for cancer genomics data. *Database* **2011:** bar026.

Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338:** 1622–1626.



# Cross-Species Analysis of Mouse and Human Cancer Genomes

Carla Daniela Robles-Espinoza and David J. Adams

*Cold Spring Harb Protoc;* doi: 10.1101/pdb.top078824; published online September 30, 2013