Published in final edited form as:

Adv Exp Med Biol. 2016 ; 916: 125–145. doi:10.1007/978-3-319-30654-4_6.

Genomic Approaches to Zebrafish Cancer

Richard M. White, M.D., Ph.D.

Memorial Sloan Kettering Cancer Center, 415 East 68th Street, New York, NY 10065, USA

Abstract

The zebrafish has emerged as an important model for studying cancer biology. Identification of DNA, RNA and chromatin abnormalities can give profound insight into the mechanisms of tumorigenesis and there are many techniques for analyzing the genomes of these tumors. Here, I present an overview of the available technologies for analyzing tumor genomes in the zebrafish, including array based methods as well as next-generation sequencing technologies. I also discuss the ways in which zebrafish tumor genomes can be compared to human genomes using cross-species oncogenomics, which act to filter genomic noise and ultimately uncover central drivers of malignancy. Finally, I discuss downstream analytic tools, including network analysis, that can help to organize the alterations into coherent biological frameworks that can then be investigated further.

Keywords

Zebrafish; Cancer; Next-generation sequencing; RNA-seq; Oncogenomics

An Introduction to Zebrafish Cancer Models

The past decade has seen an explosion in the number of available zebrafish cancer models [1]. These range from transgenic overexpression of dominant acting oncogenes, to inactivation of tumor suppressor genes, and carcinogen induced tumors. One of the major advantages of performing cancer studies in zebrafish is that they can be easily manipulated with genetic tools, and the advent of CRISPR methods [2] will only continue to accelerate this process. In addition, the optical clarity of the developing larvae or adult *casper* strain [3] allows for in vivo imaging studies that would be prohibitive in typical murine models.

Regardless of the mode of oncogenesis, all of these tumors recapitulate certain aspects of human tumorigenesis in much the same way that mouse models do, albeit at a greater speed and with a larger number of available animals to study. For example, zebrafish models of BRAF-driven melanoma strongly resemble human melanoma at the histological level [4], and in some ways, more so than mouse models do because zebrafish melanocytes are embedded in the dermal-epidermal junction like they are in humans. But in other respects, zebrafish models of cancer are divergent from human tumors, in the sense that they are generally less aggressive than many human cancers once metastasized. This may be in part

due to the fact that most fish cancers are initiated with only a few genes, whereas most human cancers harbor hundreds to thousands of mutations, copy number changes, and structural rearrangements [5].

In order to continue improving the existing fish models, we must develop methods for interrogating the genomes of these tumors to discern where the similarities, and differences, occur compared to human tumors. Such information will be invaluable in taking full advantage of the genetic and optical strengths of the zebrafish system in a manner that complements what is available in murine, fly and human cancer models. The purpose of this review is to discuss methods for genomic analysis of zebrafish tumors, with a particular eye towards a comparison to human cancer genomics.

Cancer Genomics: A Primer

The term “cancer genomics” has evolved to mean any approach which uses large-scale methods to interrogate DNA, RNA, protein, chromatin or other molecules within tumor tissue. In its earliest iterations, technologies such as PCR and Sanger sequencing were used to analyze protein coding mutations found in the exons of pancreatic cancer [6]. With the advent of “next-generation” technologies such as the Illumina HiSeq platform (see below for more details), the ability to query large numbers of tumor genomes has become feasible from both a time and cost standpoint, and has led to a rapid increase in the number of such studies.

Because of the recognition that producing high quality genomic analyses of many tumors is complex and requires a great deal of expertise, several consortia have been formed to enable quality control and scalability. In the United States, this has taken the form of The Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov/>) which has aimed to characterize several thousand tumor-normal genome pairs from nearly every malignancy. The TCGA is strongly complemented, and in some cases subsumed, by international efforts such as the International Cancer Genome Consortium (ICGC, <https://icgc.org/>) and the Cancer Genome Project at the Wellcome-Trust Sanger Center (<https://www.sanger.ac.uk/research/projects/cancergenome/>). Together, these and other efforts have begun to yield conclusive genomic data for most cancers, and publications documenting their successes are beginning to appear in the literature.

The initial phase of the TCGA effort began in 2005, and focused on glioblastoma, lung and ovarian cancer, which rapidly moved into phase II in 2009 for other tumors. The vast majority of these early efforts centered on whole-exome sequencing, in order to identify recurrent or unique mutations in the protein coding regions. However, more recently, at least three other technologies have been incorporated into the TCGA effort: array CGH (typically using Affymetrix SNP 6.0 chips), RNA-seq (both mRNA and miRNA) as well as reverse phase protein array (RPPA). Together, these datasets for a given individual tumor are analyzed using so-called “integrative landscape” analyses, in order to try and give a picture of the entire array of molecular changes present in a patient’s tumor.

The responsibility for each piece of the project is spread out across different centers, including those responsible for tumor acquisition, sequencing itself, informatics, analysis and downstream studies. For this reason, within each tumor type, as the data from each of these subprojects is collected, there is an embargo placed on the publication of the data until all members agree to a final, consortium publication from the TCGA. An illustrative example of such an approach is provided by the recent TCGA-sponsored publication for cutaneous melanoma [7]. In this manuscript, they describe an integrated DNA, RNA and protein analysis of 333 tumors, which were derived from 331 patients. The majority of these samples came from “regional” lymph node metastases, which are typically the most accessible to clinicians. Such accessibility is an important caveat for all large-scale sequencing studies in humans, since investigators are largely limited to what is practically available. For 199 of the samples, they were able to complete six types of analysis: whole-exome sequencing, DNA copy-number profiling, mRNA-sequencing, miRNA sequencing, DNA methylation profiling, and RPPA protein analysis. They also selectively used whole-genome sequencing and locus-specific PCR/Sanger sequencing of the TERT promoter for some samples. They then used an algorithm called iCluster [8], a platform for integrating disparate datasets to define the molecular subtypes of tumors, along with integrated pathway analysis to identify the major molecular alterations present in each subtype. From this, they could identify four major genomic subtypes of melanoma (BRAF, RAS, NF1, triple-negative), along with a transcriptomic subclassification into immune, keratin or MITF-low types. Interestingly, this analysis was also coupled with survival data on the patients, which implicated a new subtype of melanoma, based on immune markers, with a better prognosis. Perhaps most importantly, this landmark study also provides numerous new pathways and genes to explore in subsequent studies, some of which (e.g. KIT) may be therapeutically targetable with existing drugs. Outside of melanoma, the collection of TCGA “Pan-Cancer” datasets [9] allows for a broad sweep of mutational processes across tumor types, which revealed a core set of 127 genes that were altered across tumor types, spanning from well-known pathways like Wnt, to novel ones like histone modifications, splicing and metabolism. The number of such analyses is growing every month.

One particular challenge that many of these studies have highlighted are the remarkably diverse ways that tumors can genomically achieve the “hallmarks” of cancer such as avoidance of apoptosis. It is apparent that while major “drivers” such as RAS and BRAF are key for many tumors, there is an increasingly “long tail” of genomic alterations which may endow tumor cells with competitive advantages. In the classic definition of significant genes in cancer, most analyses have defined “driver” events as those that occur recurrently across tumors from different patients, and those falling below that cutoff as “passengers”. Of course, this is somewhat arbitrary and based on statistical likelihood, taking into account the overall mutational burdens/patterns in a given tumor type. But it is also possible that many genes deemed a “passenger” at the population level may indeed be a driver in that individual patient. Adding to this complexity is the recent finding that subclones that make up a small percentage of a given tumor (e.g. those expressing IL11) can have dramatic cell non-autonomous effects on tumor growth [10]—in a typical genomic analysis, these would likely fall below thresholds for significance, yet they are essential for the overall success of the tumor. In addition, the extremely high mutational burden of some tumors makes it extremely

difficult to assess which of those changes are truly important for tumor growth. Examples of these include UV-related melanomas, smoking-related lung cancers, as well as tumors deficient for DNA proofreading mechanisms such as those with POLD or POLE mutations [11].

It is in unraveling the complexity of human cancer genomics that the zebrafish may prove the most useful. In its simplest iteration, one can envision using a “cross-species” oncogenomics approach, in which the genomic landscape of a zebrafish cancer can be directly compared to that of the corresponding human tumor to find things in common between the two. In essence, this is a filtering approach, with the logic being that any alteration present in both species is likely to be a true driver event. It is designed to narrow down large potential genomic alterations into something manageable that can then be tested using focused downstream experiments. Other variations of this theme are cases in which lists of candidate alterations from human tumors can be tested, singly or in combination, using transgenic fish models [12]. All of these approaches will be discussed in more detail below, but it is first important to understand what has been done so far in zebrafish cancer genomics, and what tools are needed for such studies.

Zebrafish Cancer Genomics

A number of studies have now used a smattering of these technologies to interrogate zebrafish tumor models. Most of the published methods have relied upon chip-based technologies, but an increasing number are now using next-generation sequencing. It is likely that nearly all future studies will use these newer technologies.

Chip-Based Approaches

Array CGH

Array-based comparative genomic hybridization (aCGH) is a method where two different fluorescently labelled DNA samples can be hybridized to a chip that has a large array of complementary DNA fragments [13]. The labelled DNA will bind to its cognate target sequences present on the chip, and everything else washed away. Because it is generally done as a competition between two sources of DNA (i.e. tumor and normal DNA), each spot on the chip can then be scanned to calculate the relative fluorescence signal between the two samples. What is spotted onto the chip is up to each individual user, but usually it is a portion of the genome either from PCR arrays, BAC clones, cDNA clones, or in some cases small fragments of the entire genome. The level of resolution of aCGH depends entirely upon how much of the genome is spotted onto the chip, and how big each fragment is. aCGH is the gold standard for identify copy number alterations (CNAs) in tumor samples. More recently, investigators have achieved the same or better resolution that aCGH using SNP arrays, in which the fragments of DNA on the chip represent hotspots of single nucleotide polymorphisms—because those are common in the genome, they give an overall reasonable view of copy number changes, albeit not quite as a very dense aCGH chip. Although both aCGH and custom SNP arrays are available for the zebrafish genome, the majority of studies thus far have utilized the aCGH platform.

Using a BAC aCGH [14], three types of zebrafish tumors were analyzed for copy number changes: melanoma, rhabdomyosarcoma, and T-cell leukemia. Several areas of common, recurrent amplification and deletion were seen across the three tumor types, but several unique abnormalities, ranging from 1 to 28 copy number changes, were seen for individual tumors of each type. For the melanoma samples, five particular BACs were seen to be amplified in half the samples, suggesting some degree of positive selection for this region. Because this BAC array did not have very high resolution, it was not possible to discern which genes were specifically amplified in these samples, but several potentially important genes are contained within the region spanned, including EP300, PIM3, COL4A2, KIT, MITF and BRAF. Similar reports have utilized aCGH techniques for malignant peripheral nerve sheath tumors (MPNSTs) [15, 16], KRAS driven rhabdomyosarcoma [17], and T-cell ALL [18].

RNA Microarrays

Conceptually similar to aCGH, chip technologies have been also extensively applied to analysis of RNA. In this case, RNA is reverse transcribed to yield fluorescently labelled cDNA fragments, which are then hybridized to a chip containing complementary sequences. The intensity of the fluorescent spot can be inferred to be proportional to the amount of RNA species present in the original sample. Most studies tend to enrich the input RNA for mRNA, using either polyA priming or ribodepletion methods to eliminate ribosomal RNAs which make up the vast bulk of total RNA. A key aspect of this technology is what spots are placed on the chip. In the most widely used platform, the Affymetrix Zebrafish 1.0 chip, this primarily consisted of cDNAs that were derived from expressed sequence tag (EST) libraries. This chip has 14,900 transcripts, which covers much but certainly not all of the zebrafish transcriptome, which is estimated at over 25,000 genes. Other arrays are available, including an updated one from Affymetrix (although not commercially available) and an Agilent array containing 21,000 (v1) to 43,000 (v2–3) transcripts. One issue with all arrays, which is not unique to any particular technology, are the limitations in annotations of the transcripts. In part, this is related to the fact that some zebrafish transcripts do not have orthologues in other, especially mammalian, species. In other cases, the annotation difficulties are related to the genome-wide duplication that occurred in teleost evolution, making it difficult to map short oligonucleotide with confidence to a particular genome region.

Despite these limitations, microarrays have been the most widely used genomic technology in zebrafish cancer. One of the earliest attempts at this was using carcinogen-induced models of hepatocellular carcinoma in fish [19], in which they compared the tumor transcriptome to that of normal liver. This revealed a surprisingly large set of dysregulated transcripts—over 2300—which corresponded to 1920 human orthologs. The authors then used the technique of Gene Set Enrichment Analysis (GSEA) to compare the zebrafish HCC signature to human cancer [20]. GSEA is an extremely important technique for performing cross-species genomic comparisons. In short, it is a statistical method that takes in two sets of data. First, the user provides all of the expression data, for all genes and all samples, for a given genomic dataset (i.e. tumor versus normal RNA). Second, the user provides a list of genes that represents a state they might be interested in (i.e. genes associated with Wnt signaling).

GSEA then uses a ranking algorithm to determine if members of a given gene set occur near the top or bottom ranking of a dataset, providing an Enrichment Score and associated p-values. GSEA is perhaps the most powerful statistical tool available to determine if fish genes are similarly enriched in human cancer genesets, especially when it is combined with the massive database of genesets available from the MSigDB database. MSigDB is essentially a collection of curated genesets representing thousands of phenotypic states, cancer and otherwise. Using GSEA, the authors were able to show that zebrafish HCC is not only enriched in human cancer in general, but most strongly in human HCC. The particular genes that formed this enrichment between the two species belonged to the Wnt/beta-catenin pathway and MAP kinase pathway. The implication of their finding is that these two pathways are so central to the biology of HCC that they are conserved across cancers that arise in species separated by millions of years of vertebrate evolution. As both of these pathways are under intense investigation for therapeutic targeting (i.e. IWR1 for Wnt, trametinib for MAP kinase), these types of studies can lead to meaningful translational outcomes.

A similar approach was taken for a KRAS-induced rhabdomyosarcoma (RMS) model [21]. Unexpectedly, expression of KRAS under the rag2 promoter led to muscle tumors in the fish. These tumors were then profiled using Affymetrix arrays (compared to normal muscle) and compared to human RMS subtypes, which revealed that the fish tumors were more similar to the embryonal, but not alveolar, types of human RMS. This is of key importance, since previous studies had not yet linked RMS to RAS activation, although this is now widely recognized to be the case in human disease. Interestingly their data also pointed out a core “RAS” signature that was not confined to RMS, but also found in pancreatic ductal adenocarcinoma, a truly RAS driven tumor. These data point out how the fish may yield unexpected pathway alterations that have strong relevance to the human counterpart, even in some cases before human genomics have made that apparent. Similar types of analyses have been performed for other induced models of HCC, including a KRASV12 [22], xmrk [23] and RAF [24].

Our group has used comparative profiling of zebrafish and human melanoma to identify specific developmental signatures common to the two species [25, 26]. Human melanoma sequencing has revealed two dominant genetic events: BRAF and NRAS mutations. Both of these have been used to create transgenic melanoma models, in which the mutated human gene (e.g. BRAF^{V600E} or NRAS^{Q61K}) is driven under the melanocyte specific mitfa promoter [27]. In the context of a p53^{-/-} background [28], all of the animals develop easily visible tumors, which were then profiled again using Affymetrix arrays. Two types of analysis were done with these datasets. First, we compared the fish tumors to human melanoma, nevi and normal skin, which showed a striking conservation of genes expressed in both species. Taking this a step further, we then asked which of the genes contained in the melanoma signature were enriched in genes normally expressed during neural crest or melanocyte development, since melanoma is known to be a “lineage addicted” cancer [29, 30]. We obtained the list of neural crest genes using the publically available ZFIN server (<http://www.zfin.org>), which provides a rich dataset of tissue specific expression during multiple stages of development. This analysis revealed a strong enrichment of the neural crest geneset in both fish and human melanomas, providing a rationale for a subsequent

chemical screen we performed to identify small molecule suppressors of this neural crest signature. This screen ultimately identified leflunomide, a small molecule inhibitor of the metabolic enzyme dihydroorotate dehydrogenase (DHODH), which acts to suppress transcriptional elongation of neural crest genes and is now being tested as a therapeutic in human melanoma [25].

Next-Generation Sequencing Approaches

Although chip-based approaches will continue to play a role in all forms of cancer genomics, especially in regards to copy number changes, it is clear that the vast bulk of data for the foreseeable will be generated using 2nd or 3rd generation sequencing platforms. Until fairly recently, most DNA or RNA sequencing was confined to a relatively small number of samples, in which a PCR step or other form of DNA isolation was performed, followed by Sanger dideoxy chain termination sequencing. This technology generally produces short sequencing runs of 300–700 basepairs, and can be automated as found in instruments such as the ABI3730, a workhorse of many early sequencing projects. This and related technologies were used to perform the initial drafts of the human genome using shotgun approaches [31, 32]. A somewhat heroic effort using similar technology provided the first draft of a human cancer genome. Vogelstein and colleagues PCR amplified all the exons from a series of breast and colorectal cancers [33, 34] and then used the ABI3730 sequencers to delineate all of the coding mutations in these tumors.

By the early 2000s, it was clear that much higher throughput sequencing technologies were need not only for cancer genomes, but for genomics in general. A discussion of the evolution of sequencing technology is beyond the scope of this review, but summaries of this can be found elsewhere [35]. In short, most modern cancer genomic studies have begun to use so-called “massively parallel”, short-fragment sequencing as typified by machines such as the Illumina GAI/HiSeq/MiSeq, SOLID and Roche 454 platforms. Although each individual piece of DNA sequenced is only 50–500 bp long, the machine can generate millions of these “reads” in parallel and relatively rapidly, which allows for near complete coverage of a genome in about a day. These technologies were brought to bear in cancer genomics by the Wellcome-Trust Sanger Center, who published the first “whole-genome sequence” of a human cancer in 2010 [36]. Using the Illumina GAI platform, Stratton and colleagues sequenced a human melanoma cell line along with matched normal cells (COLO829 and COLO829BL). In the melanoma, they identified over 33,000 somatic mutations, along with ~900 insertion/deletion events and 51 structural rearrangements. Only a small subset (292 of 33,345 mutations) were found in protein coding regions, highlighting that the exome sequencing approach will only capture a very limited landscape of cancer genomic changes. Since that time, thousands of human cancers have been sequenced either by exome sequencing or by whole-genome sequencing using the Illumina and related platforms. More recently, so-called “third generation” sequencing technologies have come on board, including platforms from PacBIO, Oxford Nanopore and IonTorrent. The major advantage of these newer systems are the dramatically longer read lengths, which can range up to 10,000 bases or more. This will massively improve the throughput and accuracy of genomic efforts not only in human cancer, but especially in model organisms such as zebrafish where genome alignment of small read fragments still remains a computational challenge.

Exome Sequencing of Zebrafish Cancers

In order to understand how zebrafish cancers compare to human cancers, our group undertook an effort to perform a large scale exome sequencing project in collaboration with the Sanger Institute [37]. We used melanoma as a prototype, because it had a particular advantage in terms of cross-species comparisons. In human melanoma, the mutational burden is very large, as mentioned above, primarily because of the high background rate of UV damage. It is believed that most of those mutations are of little functional consequence. In contrast, transgenically engineered zebrafish melanomas have essentially no UV exposure, allowing for a direct comparison between the two species to find the most likely true drivers. In this sense, the fish mutations act as a “filter” on the human mutations.

We performed whole-exome capture on a series of 53 transgenic zebrafish melanomas, along with matched normal tissue from that animal. Most of the tumors were of the *mitf*-*BRAF*^{V600E};*p53*^{-/-} variety, with the rest being *mitf*-*NRAS*^{Q61K};*p53*^{-/-}. Several of the fish had additional candidate “driver” events built into the transgenic (e.g. *SETDB1*), in order to determine how increasing complexity of the transgenic affected the ultimate tumor genome. For each animal, tumor and normal DNA were isolated, and the exonic DNA was captured using the Zebrafish Agilent All Exon SureSelect technology. The captured DNA was sequenced using a variety of next-generation technologies, including the Illumina GAIIX, HiSeq and Roche 454 platforms. The sequence reads were mapped to the Zv9 reference genome using the standard Burrows-Wheeler Algorithm. Several types of analyses were then performed: (1) mutations were called using the CaVEMan, SomaticSniper, and String Graph Assembler algorithms, (2) insertion/deletions (indels) identified through Pindel, (3) copy number variants (CNVs) were called using the ASCAT algorithm. Because the zebrafish genome has a relatively high number of germline SNPs (compared to humans), it is absolutely essential that normal DNA is sequenced alongside tumor for all zebrafish studies. Otherwise, simply using the zebrafish reference genome as the determinant of “normal” is fraught with problems and will give an exceptionally high false positive rate.

From the 53 tumors, a total of 403 point mutations were identified, along with 13 indels and 991 copy number variants. It is striking that, on average, a median of four exonic mutations per tumor was found. This is in stark contrast to UV-related cutaneous human melanoma, which has a median of 171 coding point mutations. However, the fish melanomas are much more in line with non-UV acral human melanomas, which have a median of 9 coding mutations per tumor. The predominant mutational signature were C > T transitions, the same that is found in UV-induced melanomas, which may indicate an underlying process favoring this substitution in melanoma, even in the absence of UV exposure. In at least one tumor, there were microclusters of mutations reminiscent of “kataegis”, a phenomenon seen in human cancer yielding localized regions of hypermutation thought to arise from a single event [38]. Interestingly, few if any of the mutations in the fish were recurrent across individual fish, suggesting that either they are not important driver events, or that each fish harbors its own unique set of drivers. For the copy number changes, there were more consistent recurrent events, particularly a large 175 kb amplicon on chromosome 3 which occurred in 10/53 tumors. This region has several potentially interesting genes (i.e. *prkascaa*, *samd1*), several of which are being followed up as potentially important genes in melanoma.

One particularly striking outcome from this experiment was the tremendous degree of heterogeneity between the tumors. The vast genetic heterogeneity of human cancer is increasingly recognized [39–41], which confounds many analyses of which changes are important across populations but also within that individual patient. Conceptually, we had envisioned that transgenically engineered fish in reasonably homogeneous genetic backgrounds would harbor much less somatic variation between animals, but the exact opposite was found. For example, although the median mutation burden was four exonic mutations per tumor, the range of mutations varied from 0 to over 40. In fact, over half of the total mutations found in the entire study were identified in just 8 of the 53 tumors. Although copy number changes did show some degree of recurrence, here too there was tremendous heterogeneity. The reasons for this heterogeneity remain unclear, but a clue to this may be found in the observation that there is an inverse relationship between the number of “initiating” drivers and the subsequent number of somatic events. In other words—the more transgenes you start with, the fewer subsequent mutations you ultimately find in the tumor. It may also be due to the possibility that fish tumors are more driven by copy number, rather than mutational, events. Considering both the human and zebrafish data, it is clear that heterogeneity in cancer has a complex underpinning, the mechanisms of which remain to be identified.

From these results, it is fair to ask whether genomic sequencing of zebrafish cancer is justified given the expense and computational resources required. The answer to this depends in part on what the goals of such projects are. From a basic, mechanistic standpoint, it is likely that deeper interrogations of mutational process in tumors from fish and humans will help us to understand how specific mechanisms of genome integrity impact tumorigenesis. For example, a cross-species comparison of the relative impact of mutations versus copy number variants could be readily approached in the fish, and yield answers to why in humans, mutation and copy number variation seem to be somewhat mutually exclusive mechanisms of tumorigenesis [42]. Other questions that can be uniquely addressed in fish cancer genomics are mechanisms of processes like kataegis, which has been suggested to be due to AID-type events. Perhaps the most interesting way that the fish models can be used is to try and gain an understanding of where tumor heterogeneity comes from. Although this was initially a somewhat surprising finding from our study, it also points out that we truly do not understand the origins of genetic heterogeneity and how it relates to tumor progression. Transgenic and CRISPR models could help reveal the underlying mechanisms of these poorly understood events. Whether the fish can be used for more translational, actionable, “filtered” list of genes remains to be determined, but the data thus far indicates that for these more clinically relevant questions, the fish can be used to: (1) model potential lists of candidate DNA mutations/CNVs that arise from human TCGA data, as has been described using the miniCoopR system [12, 43], or (2) identify conserved pathways across species using RNA-seq approaches.

RNA-Sequencing of Zebrafish Cancer

As mentioned above, the vast majority of transcriptional profiling of zebrafish tumors has been done using microarray, chip based technologies. But as 2nd and 3rd generation sequencing technologies come down in price, and the informatics become more

straightforward, it is likely that nearly all such studies will migrate to RNA-seq in the near future.

Our own group has used the zebrafish BRAF^{V600E} melanomas to compare the performance of RNA-seq vs. the Affymetrix array platform (unpublished observations). In this experiment, we took total RNA and then split it to be used for either technology. The preparation for RNA-seq involves enrichment of mRNA, depletion of ribosomal RNA (very important for RNA experiments), reverse transcription into cDNA, and then fragmenting of the cDNA into small fragments. Adapters for a given sequencing platform are then ligated (in our case, the Illumina HiSeq2000 platform) in preparation for a sequencing “run”. For most applications, read lengths of 50 bp is adequate, although the longer 100 bp or more reads will improve mapping (see below for more details on this). The sequences were then aligned using Burrows-Wheeler Algorithm, and transcript assembly and differential expression accomplished with a series of algorithms: TopHat, Cufflinks and Cuffdiff [44]. There are innumerable algorithms for doing similar types of analyses of RNA-seq data. The relative transcript abundances for each sample are calculated based on the number of “reads” aligning to a particular region of protein-coding genomic sequence, and are normalized to the total amount of RNA present in a given sample.

In general, we found that the melanoma data from the two approaches was very similar, especially for the more well-expressed genes. One major difference between the technologies is the vastly greater amount of information that can be extracted from the RNA-seq. In the chip-based approaches, the most you can generally learn about a gene is the expression of a fragment (usually a 3' EST) of the gene. In some more advanced chips, which contain all of the exons, you can discern information about alternative splicing/exon usage. But in RNA-seq, this data about splicing is immediately apparent and a standard part of the analytic pipelines, because in this case one will receive all of the information about which part of a transcript is actually being expressed. This can actually cause some degree of computational headaches, because it can make it challenging to come up with a single number that represents the “expression” level of a given gene. In other words, is it better to “average” the total number of reads for all exons in a given transcript, or is it better to “sum” the total number of reads for all exons in that transcript? For situations in which your two samples use the same exons, either approach is fine, but in cases where one sample strongly uses one exon over another, this can cause a misrepresentation of the actual biology. The analysis of alternative splicing is a major advantage of RNA-seq compared to microarrays. Other more recent RNA-seq advances have included the capacity for calling underlying DNA mutations (inferred from the RNA reads), RNA editing events, strand-specific transcription, microRNA profiling, and long noncoding RNA analyses (lncRNA).

A recent study from Gong [23] used RNA-seq to analyze hepatocellular carcinomas that arise as a result of transgenic expression of xmrk in the liver. They also profiled the tumors during regression phases. This revealed a similar expression pattern to that seen in a human subtype of HCC (S2) characterized by high Myc expression and enhanced proteasome, antigen processing, p53 and cell cycle pathways. The signature of the regressing tumors was characteristic of an immune response, which is typical of EGFR expression seen in human HCC treatment.

A Practical Guide to Zebrafish Cancer Genomics: The Nuts and Bolts

Pre-experiment Planning: The Most Important Step

Because of the rapidity with which many of the above discussed technologies have moved, it is important to determine when, if and how to use them for analyzing a zebrafish tumor model. In my experience, the single most important factor in a zebrafish cancer genomics project is planning the experiments before the money gets spent. This involves three people: (1) the person doing the lab work on the zebrafish tumor, who will (hopefully) know what genomic question they are interested in, (2) a person from the genomics facility who will be doing the actual sequencing, and (3) a bio-informatician capable of analyzing the resulting data. I cannot emphasize enough how important it is to have a conversation with #2 and #3 long before a pipette is picked up. The genomics facility will almost always have key information about a technology you are interested in, exactly what your samples need to look like, how to assess quality of your samples, what type/number of “reads” you will need to do, what machine should be used for the sequencing, should you multiplex your samples, the effect of batch variation, and the file formats they are capable of producing. Similarly, the informatician is essential for telling you how “deep” your coverage should be for your given question, is it better to do single end vs. paired end reads, what machines are going to be better than others, stranded vs. nonstranded reads, how to assess quality of the genomics facility output, how are they going to be able to access the giga or terabytes of data that you produce, what types of algorithms they feel comfortable running, are they willing to develop custom algorithms, do they feel comfortable working with zebrafish (as opposed to mammalian) datasets. Together, the genomics core and informatician will provide you with the most important framework for understanding if your experiment is likely to work, how much it will cost, and how long it will take. In general, the informatics takes longer than one might assume: informaticians are in extremely high demand at this time, and often spread across many projects, and very few have experience working with the zebrafish genome. Getting this right in the beginning will have tremendous benefit down the road, as developing a good relationship with a genomics facility and an informatician will then allow subsequent projects to move much more rapidly.

Sample Preparation/Requirements

Whether you are sequencing, DNA, RNA or chromatin, one of the first questions that always comes up is “how much do I need”? There is no straightforward answer to this, in part because it depends on what you are trying to achieve. For most techniques, somewhere around 1 µg of purified nucleic is plenty, but this is probably overkill for many projects. For DNA based approaches, 1 µg will allow you to do exome or whole-genome approaches with ease, but we have also had success with as little as 100 ng (or less in many cases). RNA and ChIP can be even more forgiving, as many facilities can handle sample inputs as low as 1 ng. One issue that will come up with these low input amounts is whether to perform a round of whole-genome amplification prior to library preparation. Some of the newer amplification methods (e.g. NuGEN Ovation) produce far less bias than older methods, so for samples with very low input (i.e. less than 1 ng) it is preferable to perform amplification rather than put in a very small amount to the sequencer that is unlikely to work. As long as all the samples in a

given experiment are handled the same way (i.e. amplified or not) it is generally ok, because they will be internally controlled.

Another issue that often comes up regards library preparation. The purified DNA or RNA must be made compatible with the particular sequencer you plan to use (i.e. an Illumina library prep is not compatible with a IonTorrent prep, etc). This can involve the addition of adapters for that particular sequencer, as well as barcodes if one is multiplexing samples. Nearly all labs can prepare libraries themselves, since it involves a fairly straightforward set of molecular techniques and the protocols are well published. However, one thing we now consider is the cost-benefit ratio of doing the library preps ourselves versus within the genomics facility. Although at first glance, it may seem more cost efficient to just do it yourself (and then use for core facility to just run the sequencer), often this is not the case: many core facilities have automated equipment for library prep that reduces the per sample cost, and increases the likelihood of success. But this will clearly depend upon the facilities available at a given institution, so it is hard to generalize. One recent development has been the availability of large-scale private genomic facilities, allowing for “out-sourcing” of a great deal of the work to these highly specialized groups. For example, we have worked with the New York Genome Center (a consortium amongst many different academic labs, <http://www.nygenome.org/>) and found this to be a very efficient and cost-effective way to complete projects. Innumerable other private companies (i.e. Genewiz, Axseq, Illumina themselves) allow you to send your samples to them and they are highly expert at library prep. In all cases, it is essential to perform some type of Quality Control (QC) step before and after preparing libraries. An initial first step is to analyze your sample on a Nanodrop or Qubit type of device, to get a sense of concentration and integrity. Beyond that, many facilities will also run devices like the Bioanalyzer (for RNA) to check things like ribosomal RNA bands (as a reflection of total RNA integrity).

How Much Sequencing?

An important consideration in these approaches is how “deep” do you want to sequence, meaning how much coverage of my sample do I want? One way of making this calculation is to determine what “X” coverage you want. This is a common nomenclature in the genomics field, which basically works out to “how many times do I want a given basepair of DNA to be read by the sequencer”. So, if we state “this sample was sequenced to a depth of 40X”, that simply means that, on average, each basepair of interest was covered by at least 40 “reads”. This is clearly an average: because the DNA on the sequencer is generally randomly fragmented, a given read may cover a particular segment of the genome better or worse than another. Some regions will wind up covered at 100X, and some at 10X, and some at 0X. But the more sequencing you do, the deeper your average coverage will become.

An example is illustrative. Let’s say you wanted to sequence the entire genome of a given cancer sample to 40X coverage. The fish genome is approximate 1.4Gb of DNA, so to cover every basepair at 40X, you would need $(1.4 \times 10^9 \times 40) = 56 \times 10^9$ bp of sequence. If your sequencing facility is going to generate 100 bp, paired-end reads (so each read generates 200 bp of usable sequence), then the total number of “reads” you would need, theoretically = $(56 \times 10^9 \text{ bp}) / (200 \text{ bp}) = 280 \times 10^6$ reads. But in reality you need probably 2–3X that amount,

because many reads will not properly align to the genome (either because of errors or because of contamination with microbial constituents of most tissues) and not all reads will be of sufficient quality to be usable. In our own group, we recently performed whole-genome sequencing on a tumor:normal matched pair, and to generate ~40X coverage for each required about 750×10^6 reads per sample, or a total of 1500×10^6 reads for the two samples. Given the current capacities of Illumina 2500–4000 series machines, which generate somewhere between $200\text{--}350 \times 10^6$ reads per lane, this would require about six to eight Illumina lanes.

It is for these reasons that exome-seq or RNA-seq are much more common than whole-genome approaches. Because the protein coding genes only make up ~2 % of the genome [45], that means to get 40X coverage of the coding genes would only require 2 % as much sequencing. On a practical level, exome sequencing is typically done to much greater depths than 40X, often in the range of 100X to 1000X. This is because the main function of exome sequencing is to identify, with high confidence, mutations, and that often requires sequencing depths to greater than 40X. So, for things like mutation calling, there is a trade-off: whole-genome sequencing gives you a broader swath of the genome, but at lower depth, so you can only confidently call the most highly significant mutations. Exome sequencing gives you a small amount of the genome, but because you can go much more deeply, it is easier to call mutations that even occur at low allelic frequencies. For RNA-seq, the goal is usually not mutation calling—most experiments are done simply for differential expression. In that case, usually $20\text{--}40 \times 10^6$ reads per sample is adequate. Given a per lane capacity of 200–300 million reads, that makes multiplexing of RNA-seq samples extremely cost-effective compared to microarrays.

Basic Informatics: Raw Data to Primary Outputs

As mentioned above, most zebrafish cancer genomics projects will strongly benefit from an experienced bio-informatician, who can implement the algorithms typically used for analyzing next-generation data. Recently, there has been a community effort to make these tools more accessible to biomedical researchers through the Galaxy web portal (<https://galaxyproject.org/>) or Genepattern server (<http://www.broadinstitute.org/cancer/software/genepattern/>). Galaxy and Genepattern are powerful and relatively simple ways to run many of the algorithms discussed below, but it is still essential to have someone with experience act as a “check” on these analysis, since they will not give much information as to whether a given algorithm is appropriate—it will simply run it for you using an intuitive, web-based interface. A brief overview of some of the basic tools will be provided here.

The data output from the Illumina and other platforms is generally in FASTQ format, which is essentially a text file where each line contains the data from an individual “read” from the sequencer. An example is shown here:

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACT
CACAGTTT

!“(***+))%%%+)(%%%)1***+*)**55CCF>>>>CCCCCCC65

The top line contains the basepairs that were called by the machine, and the bottom line are the quality characters assigned to each of those, using a nomenclature specific to each platform. This format allows the subsequent algorithms to assign confidence to any given sequence using the quality metric, which is important for mapping them to the genome.

Once you have a collection of FASTQ files (each sample gets its own), these are then used as input into a “mapper” or “aligner”, which attempts to align each read to its appropriate place in the genome. From an informatics standpoint this is the critical step, and in the zebrafish is especially challenging due to the less-well developed genome (compared to humans), an increased number of gene duplications (which make it hard to discern where a given 50 bp read “belongs”) and an increased number of SNPs (which make it hard to discern if the read really belongs in that spot). There are many different mapping algorithms, but a common one for mapping DNA-seq reads is BWA (Burrows-Wheeler Algorithm). For mapping RNA-seq reads, a common one is the Bowtie/Tophat combination [46], since this allows not only for mapping of the sequence itself, but it also takes into account splice junctions where a given RNA read is likely to span two exons. These algorithms take in FASTQ files and produce files in the SAM/BAM format, which are essentially the sequence reads along with their genomic positions.

Following mapping of the reads, the next set of algorithms depends entirely upon what the question is. For detecting mutations in DNA, the TCGA has implemented a set of algorithms that are bundled into a pipeline called “Firehose” (<http://gdac.broadinstitute.org/>), primarily developed at the Broad Institute. Firehose itself is composed of several underlying algorithms including GATK [47] and MuTect [48], which try to determine whether a given basepair is different than the reference genome. Let us assume you have two samples—tumor and normal. You will generate FASTQ files for those, then map them to the genome to produce a BAM file of aligned reads. GATK will then take in the BAM files and determine if each basepair is different from the reference genome, and MuTect will then take in that data and determine if the tumor has a statistically different basepair than the matched normal sample. It will then generate a list of mutations in the VCF file format, which is essentially a list of what the basepair in the tumor is, compared to the normal sample along with the reference genome.

Mutation calling is not an either/or. It is a statistical argument as to whether a tumor sample has a higher likelihood of differing from the reference genome compared to the normal sample. An example: let us take a hypothetical chromosome position 3:1875678. At that position, let’s assume that both the tumor and normal generated 100X coverage (e.g. 100 reads). Now let’s say those reads are: tumor (G = 99, A = 1), normal (G = 1, A = 99) and the reference genome at that position = A. In that case, it is likely that the tumor has a mutation at that position, and the single “A” in the tumor sample is either a sequencing error or normal tissue contamination. But now let’s say the reads are: tumor (G = 80, A = 20), normal (G = 3, A = 97) and the reference genome at that position = A. Is that position mutated in the tumor, or is it actually just heterozygous or contaminated with normal DNA? MuTect and related algorithms try to determine the statistical likelihood of mutation, taking these factors into account, but outside of very clear examples where the allelic frequencies are close to 100 %, it can be a difficult judgment to make.

For detecting differential gene expression from RNA, again there are many algorithms, but some commonly used pipelines include Cuffdiff (<http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/>) and DeSeq (<http://bioconductor.org/packages/release/bioc/html/DESeq.html>). Here, the process is again to map all of the reads to the genome, and then generate a number which represents the level of expression of a given gene. The most commonly used one is FPKM, which stands for Fragments Per Kilobase of transcript per million mapped reads. Calculation of FPKM is essentially a normalized view of transcription for that given gene, and can be used to then compare across samples using statistical methods such as False Discovery Rate (FDR). Similar to a microarray analysis, the output from this analysis typically gives both a fold-change (tumor compared to normal) as well as a p-value and corrected q-value (typically from the FDR calculation). Genes can then be stratified by the user to determine what level of significance is most meaningful to them, i.e. a $q < 0.05$, a $q < 0.01$, etc. Similar to DNA-seq, there is no “either/or” for significance—it is simply a level of significance you feel comfortable with.

Downstream Analyses: Secondary Informatics Tools to Assign Meaning to the Data

All of the primary outputs from sequencing projects, whether they be DNA, RNA, or ChIP based, will ultimately generate a list of genomic fragments that are different between tumor and normal. Some of these differences will be in genes themselves, some will be in noncoding RNAs, promoters, enhancers, intergenic regions, UTRs, etc. It is tempting to initially just look at a list of these, especially ones in genes, and try to make a biological story from these lists. But this can be challenging to do when that list contains thousands of genes or regions, and so a large number of downstream tools have been developed to try and connect these findings to the broader biological literature. In fact, one can argue that individual investigators should be very cautious in looking at lists of genes and interpreting it without some type of secondary analysis, since the inherent biases we all have (based on our prior knowledge) can skew the meaning to be found in these large datasets. For this reason, it is important to use secondary tools with some statistical power to connect the data.

Pathway analysis is a common method to try and determine if your list of genes or regions are somehow connected to each other. Some commonly used ones are DAVID (<http://david.abcc.ncifcrf.gov/>), KEGG (<http://www.genome.jp/kegg/>), GO/Gene Ontology (<http://geneontology.org/>), GREAT (<http://bejerano.stanford.edu/great/public/html/>) and IPA (Ingenuity Pathway Analysis, <http://www.ingenuity.com/products/ipa>). IPA is particularly useful for zebrafish studies, since it will allow you to directly input zebrafish annotations, and then cross-reference that to data from other species on pathways. IPA is especially good at incorporating in not only the gene name, but also the level of significance for that particular gene in your dataset (i.e. q-value or fold-change). It will output a discrete set of “canonical” pathways altered in your dataset (i.e. Wnt/beta-catenin signaling or PI3K signaling, etc), and give you a corresponding p-value to determine how likely it is that your dataset is truly enriched in that pathway. IPA uses a combination of automated and manual curated pathways to connect genes to each other, and provides the level of evidence for those associations. It also includes a unique tool called “Upstream Regulatory Analysis”, which

attempts to statistically predict what upstream factors may have been responsible for a given gene expression signature (i.e. EGFR signaling might produce a given gene expression signature in lung cancer cells, etc.). The combination of pathway analysis plus upstream regulator analysis often leads to testable, discrete hypotheses that can then be tested in the laboratory. IPA a proprietary product that is continuously updated, and many academic institutions have subscriptions. Other tools like DAVID can provide somewhat similar data, and although not quite as comprehensive, it is free and very straightforward to implement.

As mentioned above, Gene Set Enrichment Analysis (GSEA) is another key tool for both pathway analysis as well as cross-species comparisons. It is a free tool provided by the Broad Institute (<http://www.broadinstitute.org/gsea/index.jsp>), and once one has mastered the unique file formats, it is very straightforward to run using its Java-based applet. The most powerful aspect of GSEA is that it is deeply connected to the MSigDB database (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>), essentially a curated list of gene expression signatures encompassing nearly all biological states of interest (i.e. a signature of lung cancer, a signature of MAP kinase pathway, etc). For this reason, you can take a given dataset that emerges from a zebrafish study, and then use GSEA to compare it to the entire MSigDB database to figure out what your sample is most similar to. GSEA provides a statistically rich output that includes p-values, false discovery rates, and a unique score called an “Enrichment Score” which provides a convenient way to gauge how similar your dataset is to another one from the literature. The GSEA applet also produces publication quality visualization tools, especially Enrichment Plots that have become widely used in the literature and are easy to interpret.

In many cases, it will be important to compare a zebrafish cancer dataset to what is known about human cancer. GSEA can be used for this, but some specific tools focusing on human cancer can be especially powerful here. The cBio Portal (<http://www.cbioportal.org/>) is a publically available tool for accessing all of the TCGA project data, which very cleanly shows all of the mutation, copy number, mRNA and phosphoproteomic data available for each tumor type. The input to the cBio Portal is human gene names, so zebrafish IDs will need to be converted to human orthologs using tools such as DRSC (http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl) or Ensembl/Biomart (<http://www.ensembl.org/biomart/>). Another useful tool for human cancer data is Oncomine (<https://www.oncomine.org/>), which has collected massive numbers of RNA expression profiles (either microarray or RNA-seq) along with a smattering of copy number and DNA-seq datasets. Oncomine is very good for querying one gene very deeply across human cancers, and determining how that gene is altered in those tumors. It is available as both a free and proprietary product (with enhanced features). A unique resource for protein data is the Human Protein Atlas (<http://www.proteinatlas.org/>), which is attempting to determine protein expression of every gene in the genome across both normal and cancerous tissues. They provide detailed data about each antibody they are using, along with validation status. They also provide photomicrographs of each sample, and a statistical measure of enrichment of a given protein in a given condition. Similar to Oncomine, it is very useful for deeply probing a particular gene for its role in human cancers.

Summary and Perspective

As the pace of studies using zebrafish as a cancer model has accelerated, so too has the pace of human cancer genomics. The major challenge over the next decade will be to determine how zebrafish cancer models can integrate with what is being done in human cancer such that the fish provides a truly unique tool in dissecting various aspects of tumorigenesis. It is nearly impossible to do this without some baseline interrogation of zebrafish cancer genomics, whether that be at the DNA, RNA, protein or epigenetic levels. The tools that have been developed for human cancer genomics can usually be applied without too much difficulty to zebrafish cancer genomics, and the number of tools that allow for cross-species oncogenomics continues to grow. The information presented here is meant to be a starting point for those interested in this interface between the two species, and what tools can be used to leverage the particular strengths of each system. These tools will continue to evolve, enabling fertile collaborations between the zebrafish and human cancer biologists.

Acknowledgments

This work was supported by the NIH Directors New Innovator Award (DP2CA186572), K08AR055368, the Melanoma Research Alliance Young Investigator Award, an AACR/ASCO Young Investigator Award, and the Alan and Sandra Gerry Metastasis Research Initiative.

References

1. White R, Rose K, Zon L. Zebrafish cancer: the state of the art and the path forward. *Nat Rev Cancer*. 2013; 13:624–636. [PubMed: 23969693]
2. Ablain J, Durand EM, Yang S, Zhou Y, Zon LI. A CRISPR/Cas9 vector system for tissue-specific gene disruption in zebrafish. *Dev Cell*. 2015; 32:756–764. [PubMed: 25752963]
3. White RM, Sessa A, Burke C, Bowman T, LeBlanc J, Ceol C, et al. Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell Stem Cell*. 2008; 2:183–189. [PubMed: 18371439]
4. Patton EE, Widlund HR, Kutok JL, Kopani KR, Amatruda JF, Murphey RD, et al. BRAF mutations are sufficient to promote nevi formation and cooperate with p53 in the genesis of melanoma. *Curr Biol*. 2005; 15:249–254. [PubMed: 15694309]
5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science*. 2013; 339:1546–1558. [PubMed: 23539594]
6. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*. 2008; 321:1801–1806. [PubMed: 18772397]
7. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell*. 2015; 161:1681–1696. [PubMed: 26091043]
8. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009; 25:2906–2912. [PubMed: 19759197]
9. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013; 502:333–339. [PubMed: 24132290]
10. Marusyk A, Tabassum DP, Altmann PM, Almendro V, Michor F, Polyak K. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*. 2014; 514:54–58. [PubMed: 25079331]
11. Heitzer E, Tomlinson I. Replicative DNA polymerase mutations in cancer. *Curr Opin Genet Dev*. 2014; 24:107–113. [PubMed: 24583393]
12. Ceol CJ, Houvras Y, Jane-Valbuena J, Bilodeau S, Orlando DA, Battisti V, et al. The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature*. 2011; 471:513–517. [PubMed: 21430779]

13. Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, et al. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nat Commun.* 2015; 6:6315. [PubMed: 25697895]
14. Freeman JL, Ceol C, Feng H, Langenau DM, Belair C, Stern HM, et al. Construction and application of a zebrafish array comparative genomic hybridization platform. *Genes Chromosomes Cancer.* 2009; 48:155–170. [PubMed: 18973135]
15. Zhang G, Hoersch S, Amsterdam A, Whittaker CA, Lees JA, Hopkins N. Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers. *Proc Natl Acad Sci U S A.* 2010; 107:16940–16945. [PubMed: 20837522]
16. Zhang G, Hoersch S, Amsterdam A, Whittaker CA, Beert E, Catchen JM, et al. Comparative oncogenomic analysis of copy number alterations in human and zebrafish tumors enables cancer driver discovery. *PLoS Genet.* 2013; 9:e1003734. [PubMed: 24009526]
17. Chen EY, Dobrinski KP, Brown KH, Clagg R, Edelman E, Ignatius MS, et al. Cross-species array comparative genomic hybridization identifies novel oncogenic events in zebrafish and human embryonal rhabdomyosarcoma. *PLoS Genet.* 2013; 9:e1003727. [PubMed: 24009521]
18. Rudner LA, Brown KH, Dobrinski KP, Bradley DF, Garcia MI, Smith AC, et al. Shared acquired genomic changes in zebrafish and human T-ALL. *Oncogene.* 2011; 30:4289–4296. [PubMed: 21552289]
19. Lam SH, Wu YL, Vega VB, Miller LD, Spitsbergen J, Tong Y, et al. Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat Biotechnol.* 2006; 24:73–75. [PubMed: 16327811]
20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005; 102:15545–15550. [PubMed: 16199517]
21. Langenau DM, Keefe MD, Storer NY, Guyon JR, Kutok JL, Le X, et al. Effects of RAS on the genesis of embryonal rhabdomyosarcoma. *Genes Dev.* 2007; 21:1382–1395. [PubMed: 17510286]
22. Nguyen AT, Emelyanov A, Koh CH, Spitsbergen JM, Lam SH, Mathavan S, et al. A high level of liver-specific expression of oncogenic Kras(V12) drives robust liver tumorigenesis in transgenic zebrafish. *Dis Model Mech.* 2011; 4:801–813. [PubMed: 21729876]
23. Li Z, Luo H, Li C, Huo X, Yan C, Huang X, et al. Transcriptomic analysis of a transgenic zebrafish hepatocellular carcinoma model reveals a prominent role of immune responses in tumour progression and regression. *Int J Cancer.* 2014; 135:1564–1573. [PubMed: 24550086]
24. He S, Krens SG, Zhan H, Gong Z, Hogendoorn PC, Spaink HP, et al. A DeltaRaf1-ER-inducible oncogenic zebrafish liver cell model identifies hepatocellular carcinoma signatures. *J Pathol.* 2011; 225:19–28. [PubMed: 21744342]
25. White RM, Cech J, Ratanasirinawoot S, Lin CY, Rahl PB, Burke CJ, et al. DHODH modulates transcriptional elongation in the neural crest and melanoma. *Nature.* 2011; 471:518–522. [PubMed: 21430780]
26. Dovey M, White RM, Zon LI. Oncogenic NRAS cooperates with p53 loss to generate melanoma in zebrafish. *Zebrafish.* 2009; 6:397–404. [PubMed: 19954345]
27. Lister JA, Robertson CP, Lepage T, Johnson SL, Raible DW. nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate. *Development.* 1999; 126:3757–3767. [PubMed: 10433906]
28. Berghmans S, Murphey RD, Wienholds E, Neuberg D, Kutok JL, Fletcher CD, et al. tp53 mutant zebrafish develop malignant peripheral nerve sheath tumors. *Proc Natl Acad Sci U S A.* 2005; 102:407–412. [PubMed: 15630097]
29. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature.* 2005; 436:117–122. [PubMed: 16001072]
30. Garraway LA, Weir BA, Zhao X, Widlund H, Beroukhi R, Berger A, et al. “Lineage addiction” in human cancer: lessons from integrated genomics. *Cold Spring Harb Symp Quant Biol.* 2005; 70:25–34. [PubMed: 16869735]
31. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science.* 2001; 291:1304–1351. [PubMed: 11181995]

32. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
33. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science*. 2006; 314:268–274. [PubMed: 16959974]
34. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. *Science*. 2007; 318:1108–1113. [PubMed: 17932254]
35. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012; 2012:251364. [PubMed: 22829749]
36. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*. 2010; 463:191–196. [PubMed: 20016485]
37. Yen J, White RM, Wedge DC, Van Loo P, de Ridder J, Capper A, et al. The genetic heterogeneity and mutational burden of engineered melanomas in zebrafish models. *Genome Biol*. 2013; 14:R113. [PubMed: 24148783]
38. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012; 149:979–993. [PubMed: 22608084]
39. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012; 366:883–892. [PubMed: 22397650]
40. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011; 472:90–94. [PubMed: 21399628]
41. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*. 2010; 467:1114–1117. [PubMed: 20981102]
42. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*. 2013; 45:1127–1133. [PubMed: 24071851]
43. Iyengar S, Houvras Y, Ceol CJ. Screening for melanoma modifiers using a zebrafish autochthonous tumor model. *J Vis Exp*. 2012:e50086. [PubMed: 23183931]
44. Kulahoglu C, Brautigam A. Quantitative transcriptome analysis using RNA-seq. *Methods Mol Biol*. 2014; 1158:71–91. [PubMed: 24792045]
45. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013; 496:498–503. [PubMed: 23594743]
46. Lindner R, Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS One*. 2012; 7:e52403. [PubMed: 23300661]
47. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20:1297–1303. [PubMed: 20644199]
48. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013; 31:213–219. [PubMed: 23396013]