PERSPECTIVE ARTICLE
published: 12 January 2015
doi: 10.3389/fgene.2014.00470

A gene mapping bottleneck in the translational route from zebrafish to human

Niek de Klein[1,2], Mark Ibberson[3], Isaac Crespo[3], Sophie Rodius[1] and Francisco Azuaje[1]*

[1] NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of Health (formerly CRP-Santé), Luxembourg, Luxembourg
[2] Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[3] Vital-IT Systems Biology Division, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
Mikhail P. Ponomarenko, Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences, Russia
Hongying (Daisy) Dai, Children's Mercy Hospital, USA
Yvonne Bradford, University of Oregon, USA

***Correspondence:**
Francisco Azuaje, NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of Health (formerly CRP-Santé), 1A-B, Rue Thomas Edison, L-1445 Strassen, Luxembourg, Luxembourg
e-mail: francisco.azuaje@lih.lu

Among a diversity of animal models of disease, the zebrafish is a promising model organism for enabling novel translational biomedical research. To fully achieve the latter, a key requirement is to match molecular readouts measured in zebrafish with information relevant to health and disease in humans. A fundamental step in this direction is to accurately map gene sequences from zebrafish to humans. Despite significant progress in genome annotation, this remains an intricate and time-consuming challenge. Here we discuss major obstacles that we had to overcome to systematically map genes from zebrafish to human. We identified important disparities, as well as partial agreements, between five public zebrafish-to-human homology resources. There is still a need for standardized, comprehensive genomic mappings between zebrafish and humans. Without this, efforts to use zebrafish as a powerful translational research tool will be stalled.

Keywords: zebrafish, zebrafish-to-human gene mapping, orthology inference, genome annotation, translational research

LAYING OUT THE ROAD

The zebrafish (*Danio rerio*) is a powerful and relatively low-cost tool for fundamental and translational biomedical research. It offers *in vivo* models with potential clinical relevance, which are valuable to elucidate disease mechanisms, novel therapeutic targets, and candidate therapeutics (Mione and Trede, 2010; Gemberling et al., 2013; White et al., 2013; Kikuchi, 2014). Substantial efforts to enable zebrafish research have been reflected in ongoing initiatives in the USA, Europe, and elsewhere[1]. Examples of recent significant outcomes include The Zebrafish Model Organism Database (ZFIN[2]; Bradford et al., 2011) and the publication of the zebrafish reference genome sequence (Howe et al., 2013). The latter estimated that about 70% of human genes have at least one unambiguous zebrafish ortholog (Howe et al., 2013).

Notwithstanding the quality and applicability of these advances, key challenges remain to translate findings from zebrafish to humans on the basis of genome-wide sequence mapping. In this translational route, we are facing a heavy bottleneck early.

THE NEED FOR ACCURATE AND COMPREHENSIVE GENE MAPPING

A crucial task for investigating translational research applications is the linking of molecular readouts measured in zebrafish to information relevant to human health and disease. To accomplish this, a key requirement is the matching of gene sequences from zebrafish to humans at a genome-wide scale. This step goes beyond the automated conversion of gene symbols, and often involves the association between multiple homologous sequences that are included in gene expression microarrays or RNA sequencing experiments.

A realistic scenario begins, for example, with the identification of a set of genes that are differentially expressed between pathological and control states, e.g., disease vs. healthy phenotypes. The resulting gene list may be mapped to homologous genes in humans using HomoloGene[3] (Wheeler et al., 2001; Acland et al., 2014) and ZFIN (Bradford et al., 2011). Additionally, researchers may require to map sequences from specific microarray platforms, for example: from Affymetrix's GeneChip Zebrafish Genome Array to GeneChip Human Genome U133A (goo.gl/d3lCPL). This is an essential prerequisite to perform search and matching of "omics" profiles related to disease and drug responses, which are stored in different databases. Prominent examples of the latter are the Gene Expression Omnibus (GEO[4]; Barrett et al., 2013), ArrayExpress[5] (Rustici et al., 2013) and the Connectivity Map (cMAP[6]; Lamb et al., 2006; Lamb, 2007).

[1] https://zfin.org/zf_info/catch/zebrafish_programs.html
[2] https://zfin.org
[3] www.ncbi.nlm.nih.gov/homologene
[4] www.ncbi.nlm.nih.gov/geo
[5] www.ebi.ac.uk/arrayexpress
[6] www.broadinstitute.org/cmap

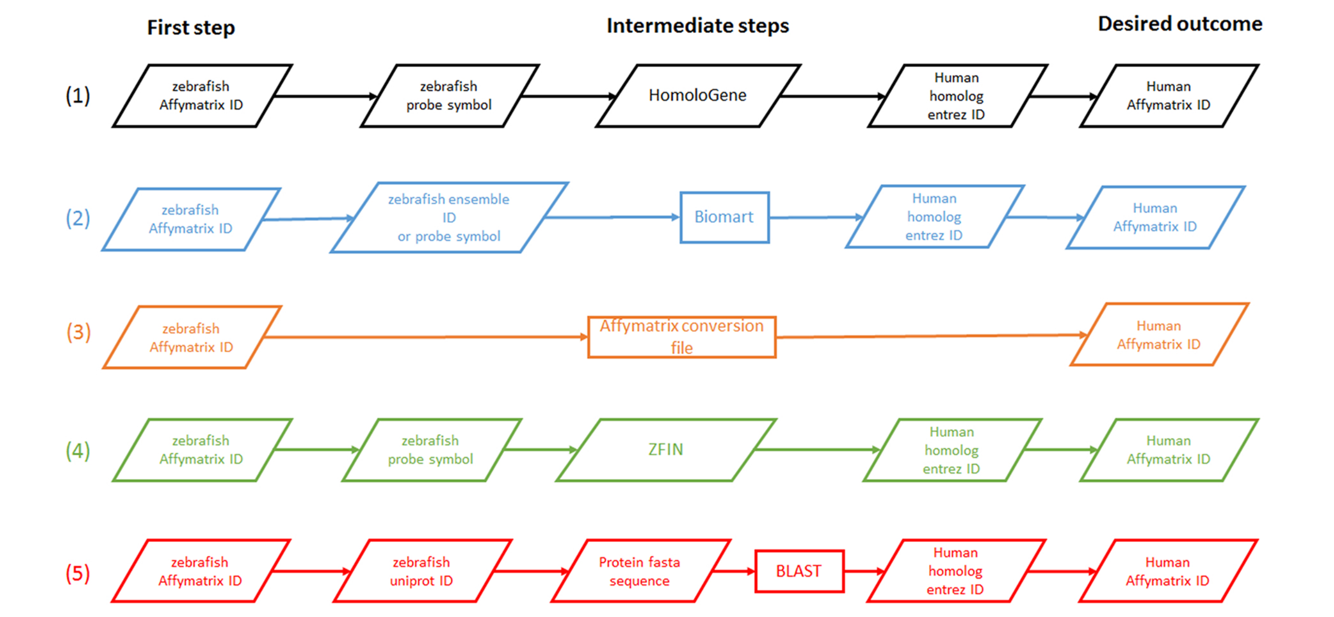

FIGURE 1 | Mapping zebrafish to human sequences via five annotated resources. Sequence probes from Affymetrix's GeneChip Zebrafish Genome Array were mapped to probes in GeneChip Human Genome U133A. Each mapping pipeline is based on a single resource independently: **(1)** HomoloGene (Wheeler et al., 2001; Acland et al., 2014), **(2)** Biomart (Kasprzyk, 2011), **(3)** conversion file provided by Affymetrix, **(4)** ZFIN (Bradford et al., 2011), and **(5)** BLAST homology searches performed at our laboratory (Altschul et al., 1990, 1997; Ye et al., 2006). In the latter method we focused on the most statistically significant BLAST match per query.

PRACTICAL CHALLENGES REMAIN

We recently implemented this process as part of a project in drug repositioning that applies the zebrafish as *in vivo* model of heart regeneration. In this particular case, we aimed to match sequence probes from zebrafish to human using the microarray chips indicated above. Initially, we expected that a single mapping resource, such as HomoloGene (Wheeler et al., 2001; Acland et al., 2014), could allow us to go from zebrafish gene symbols to human homolog symbols (e.g., Entrez database IDs) in a relatively straightforward way. After testing options available, we understood that a single resource does not provide up-to-date, comprehensive genome-scale mappings.

To overcome this obstacle, we implemented a pipeline that incorporated five gene mapping resources: (1) HomoloGene (Wheeler et al., 2001; Acland et al., 2014), (2) Biomart[7] (Kasprzyk, 2011), (3) conversion file provided by Affymetrix, (4) ZFIN (Bradford et al., 2011), and (5) BLAST homology searches performed at our laboratory (Altschul et al., 1990, 1997; Ye et al., 2006; **Figure 1**). Most of these resources (HomoloGene, ZFIN, and Biomart) apply a combination of: expert curation of orthology relationships found in the literature, manual orthology analysis, and (only) computational prediction of orthology. The other two resources are based on associations provided by the microarray manufacturer and our own computational predictions without deep expert curation. The human homolog Entrez IDs resulting from each procedure were compared, and overlaps among them were identified (**Figure 2**).

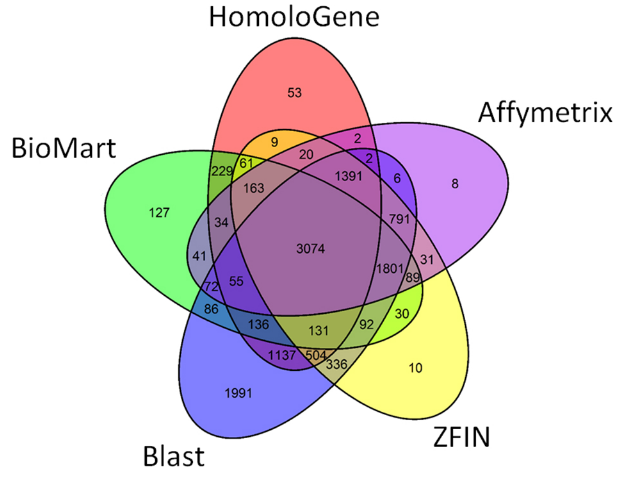

FIGURE 2 | Gene mapping agreement between the different homology annotation resources. Number of mapped genes are shown for each resource and between-resource intersection.

Using the five methods, a total of 12,593 human genes (Entrez IDs) were mapped from the 13,287 zebrafish sequence probes used as inputs (94.8%). Among the public resources, ZFIN (Bradford et al., 2011) reported the largest number of mapped genes (8533), followed by the Affymetrix conversion file (7580), HomoloGene (Wheeler et al., 2001; Acland et al., 2014) (7001), and Biomart (Kasprzyk, 2011) (6221). Our BLAST-based analysis (Altschul et al., 1990, 1997; Ye et al., 2006) generated 11,605 matches, including 1991 mappings that were not found in any of

[7] www.biomart.org

the other resources. Together, the five resources jointly agreed on 3074 mappings only (24.4%). Also there were relatively narrow overlaps between HomoloGene (Wheeler et al., 2001; Acland et al., 2014), a high-quality expert-annotated database, and the other resources (**Figure 2**). When two or more resources produced a mapping for any given zebrafish gene there was always an agreement in their mappings.

Considering these disparities and incompleteness, we decided to incorporate all available mapping evidence into the subsequent stages of our project. Although we assigned more confidence to mappings originating from expert-curated databases or to those supported by multiple resources (**Figure 2**), we also had to consider on a case-by-case basis those instances only mapped by a single source (e.g., Affymetrix file) or those only available in our BLAST search (Altschul et al., 1990, 1997; Ye et al., 2006). This semi-automatic, time-consuming conversion process was required for every zebrafish-derived candidate signature obtained in our project.

The major take-home message from **Figure 2** is that whilst a large number of orthology predictions overlap between the five resources, this only amounts to just over 24% of the total number of annotated genes. This means that a straightforward voting approach only assigning orthologs commonly assigned by all resources, would mean ignoring 75% of genes, which is unacceptable for any genome-wide experiment. Even when taking just two resources such as Homologene (Wheeler et al., 2001; Acland et al., 2014) and ZFIN (Bradford et al., 2011), 1648 genes (24%) of the Homologene mappings are not found in ZFIN. The situation is even more striking considering ZFIN, where 3186 (37%) of mappings are not found in Homologene. Thus, the interpretation of whole genome experiments from zebrafish in a human context will be strongly affected simply by the choice of resource used for the orthology mapping.

OVERCOMING OBSTACLES TO ENABLING RESEARCH

Our experience illustrates that performing zebrafish-to-human gene mapping remains a major challenge. This is a critical requirement for enabling research in different application domains. As zebrafish becomes a widely adopted genetic and systems biology model of disease, comprehensive and accurate zebrafish-to-human gene mapping represents a fundamental need. The complexity of this endeavor is magnified by intertwined evolutionary and genomics factors, including the considerable levels of gene similarity at the genome and gene family levels.

Finding maximal, high-quality sets of orthology relationships is currently constrained by the incompleteness of the zebrafish genome assembly, and in general by the evolutionary separation between species. Although our analysis considered (zebrafish) microarray probes that are linked to zebrafish genes previously annotated, this is an important factor to consider regardless of gene selection scheme or the lack of agreement between databases.

The *de novo* implementation of this process may be too time-consuming or even impractical for many laboratories, in particular those with limited bioinformatics resources. Moreover, even if bioinformatics capacity is available, the required information will continue maturing and evolving. This is very likely as the annotation of the zebrafish genome goes deeper and new evidence about gene function emerges. Furthermore, progress will be accompanied by a fast-growing interest in non-coding RNA sequences.

On one level, the agreement between gene homology resources highlight the confidence strength for such zebrafish-to-human mappings. On a gene-by-gene basis, databases that make major efforts to incorporate expert curation, ZFIN in particular, are likely to offer the highest quality relationships when those mappings are available. On another level, the considerable complementarity among these resources underlines the need for further annotation efforts, as well as their integration and standardization. Future comparisons could include other resources of orthology inference that were not considered in our analysis, such as the PANTHER classification system[8]. The incorporation of curation "evidence codes" (e.g., literature-extracted vs. manual orthology analysis) may also benefit the usage and integration of available resources. Future work could also benefit from incorporating phylogenetic evidence using multiple animal models and species.

Until a more standardized solution exists, researchers should not rely on a single resource for zebrafish ortholog mapping. Rather, we recommend using a combination of several resources and performing focused manual annotation on subsets of genes falling between annotation categories. In order to keep such numbers manageable this annotation could be restricted to small subsets of genes showing key biological relevance for the experiment in question. Ideally such annotation would be fed back into manually curated resources such as ZFIN (Bradford et al., 2011), thus making the annotations available to other researchers in the field. Researchers are welcome to request from us the multi-source mapping data discussed in this article.

The zebrafish can provide us with significant biological insights and novel directions for therapeutic interventions in a wide range of disease domains, including cardiovascular disease and cancers. To accomplish this vision, comprehensive and accurate zebrafish-to-human gene mapping is still necessary. Further public standardized efforts are needed. This will greatly depend on stronger support from research funders, researchers and other stakeholders.

ACKNOWLEDGMENTS

This research was funded in part by the INTER program of Luxembourg's National Research Fund (FNR) and the Swiss National Research Foundation (SNF), INFUSED project[9]. We also thank support from SNF (funding #145260). We thank support from G. Androsova, F. Bernardin, C. Jeanty, A. Muller, P. V. Nazarov, S. P. Niclou, and L. Vallar at CRP-Santé; N. Mercader (CNIC, Spain); and A. Feenstra (VU, Netherlands). For their helpful feedback on this contribution, we are grateful to D. Howe and K. Frazer at ZFIN (USA), and I. Xenarios at SIB (Switzerland).

REFERENCES

Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C. et al. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 42, D7–D17. doi: 10.1093/nar/gkt1146

[8] www.pantherdb.org

[9] www.infused-project.eu

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D. G., et al. (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* 39, D822–D829. doi: 10.1093/nar/gkq1077

Gemberling, M., Bailey, T. J., Hyde, D. R., and Poss, K. D. (2013). The zebrafish as a model for complex tissue regeneration. *Trends Genet.* 29, 611–620. doi: 10.1016/j.tig.2013.07.003

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi: 10.1038/nature12111

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database* (*Oxford*) 2011, bar049. doi: 10.1093/database/bar049

Kikuchi, K. (2014). Advances in understanding the mechanism of zebrafish heart regeneration. *Stem Cell Res.* 13, 542–555. doi: 10.1016/j.scr.2014.07.003

Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* 7, 54–60. doi: 10.1038/nrc2044

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Mione, M. C., and Trede, N. S. (2010). The zebrafish as a model for cancer. *Dis. Models Mech.* 3, 517–523. doi: 10.1242/dmm.004747

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., et al. (2013). ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.* 41, D987–D990. doi: 10.1093/nar/gks1174

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., et al. (2001). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 29, 11–16. doi: 10.1093/nar/29.1.11

White, R., Rose, K., and Zon, L. (2013). Zebrafish cancer: the state of the art and the path forward. *Nat. Rev. Cancer* 13, 624–636. doi: 10.1038/nrc3589

Ye, J., Mcginnis, S., and Madden, T. L. (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 34, W6–W9. doi: 10.1093/nar/gkl164

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 November 2014; accepted: 19 December 2014; published online: 12 January 2015.

Citation: de Klein N, Ibberson M, Crespo I, Rodius S and Azuaje F (2015) A gene mapping bottleneck in the translational route from zebrafish to human. Front. Genet. 5:470. doi: 10.3389/fgene.2014.00470

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.