



Improving on the identification of the Human Orthologs to Zebrafish Genes through an Extended Map

Wasay Hussain^{1,2}, Andrea Sboner^{1,2}, Mark A Rubin¹, Yariv Houvras³, Francesca Demichelis^{2,4}

Abstract # 12600517

¹ Department of Pathology and Laboratory Medicine, ² Institute for Computational Biomedicine, and ³ Department of Surgery, Weill Cornell Medical College, New York, NY USA
⁴ Centre for Integrative Biology, University of Trento, Trento, Italy



Abstract

Zebrafish is becoming an important model organism to functionally characterize the biological impact of disease specific mutations in humans. However, the translation of the findings back to the human setting is impaired in part by the incomplete characterization of the orthology between zebrafish and human genes. In fact, nearly one-fourth of zebrafish genes are not annotated to indicate an orthologous relationship with a higher vertebrate, or human gene.

To address this, we implemented a step-wise process to build a transcriptome map identifying zebrafish-human orthologs: ZFis(H)uman. First, a two-way sequence alignment step identifies and ranks pairs of zebrafish gene and human orthologous gene by tracking the alignment quality measures. Then, based upon the ranking, sequence conservation information and zebrafish transcript abundance as evaluated by RNA-seq are considered. At each step well-characterized genes are used as a baseline for statistical analyses. This approach identifies human orthologous genes with statistical features that are similar to known characterized ones. The two-way alignment step identified high quality matches for 8,445 zebrafish RefSeq genes, 21% of which are currently 'uncharacterized'. Other 2,426 genes were highly ranked based upon estimated orthologous relationship, of which 24% are 'uncharacterized'. Overall the portion of zebrafish transcriptome that could not be linked to the human transcriptome is reduced from ~25% to ~10%.

Methods

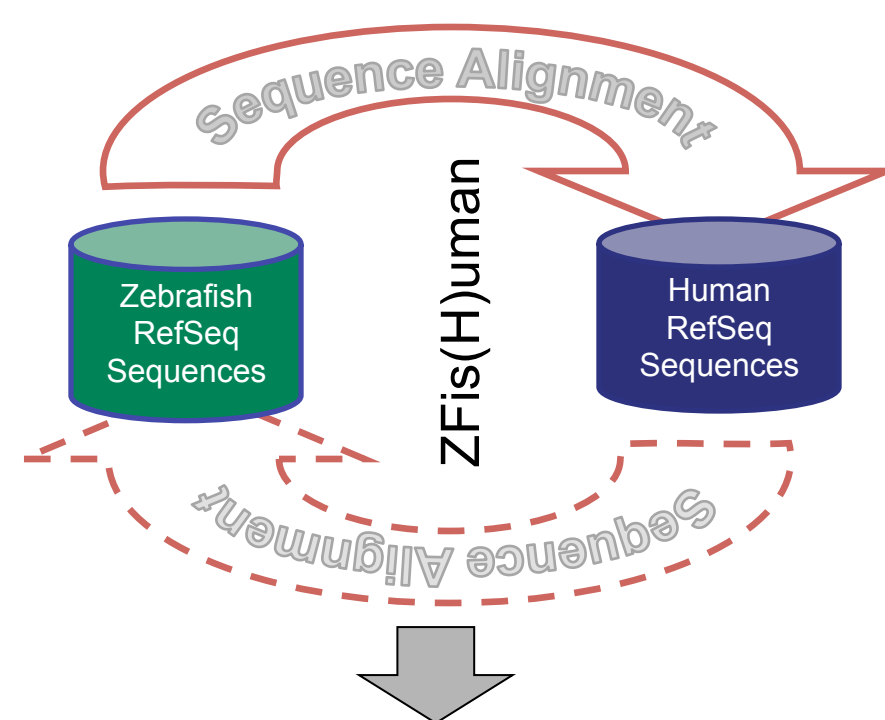


Figure 1. Schematic of the ZFis(H)uman approach. The sequences of zebrafish genes are aligned against the sequences of the human genes using blastx.

Each corresponding human gene is then aligned back to the zebrafish reference. All its hits, i.e. zebrafish genes, are considered for characterization of orthology (see below).

Considering *human to zebrafish* alignment:

- **1st Hit** = the first hit matches the zebrafish gene
- **2nd Hit** = the second hit matches the zebrafish gene
- **3rd Hit** = the third hit matches the zebrafish gene

Table 1. Characterization of human orthologs.

Zebrafish	Human hit	Zebrafish hit	Orthology Characterization
<i>fgf6</i>	FGF6	<i>fgf6</i>	1 st Hit
<i>braf</i>	BRAF	<i>braf</i>	1 st Hit
<i>zgc:101635</i>	RTKN	1:si:dkey-40c11.1 2: <i>zgc:101635</i>	2 nd Hit
<i>drd1</i>	DRD1	1:LOC7926 2:CAP1 3: <i>drd1</i>	3 rd Hit

Table 2: Statistics reported for each ortholog.

Zebrafish	Human hit	E-value	Percent identity	Length (bp)
<i>fgf6</i>	FGF6	1e-79	63.64	187
<i>braf</i>	BRAF	<1e-200	94.83	348
<i>zgc:101635</i>	RTKN	6e-98	62.07	203
<i>drd1</i>	DRD1	7e-143	76.03	146

- *E-value* = Probability of hit occurring by chance alone. The lower the E-value, the better the hit
- *Percent identity* = The identity between the query sequence and the subject hit
- *Length* = size of the subject hit

Query* = Comparing sequence **from
Subject* = Comparing the sequence **to

Zebrafish – Human Orthologs

Table 3. Summary of the zebrafish genes with human orthologs determined as described in Methods. Only a small fraction of zebrafish genes (~4%) does not have a human ortholog.

	Zebrafish	Human orthologs
Characterized	11,394 (74%)	10,746 (74%)
Un-characterized	4,025 (26%)	3,785 (26%)
Total	15,599*	14,404* (96%)

* Data downloaded from UCSC Genome Browser RefSeq (refGene) table:
• Zebrafish: Jun 29, 2011 (Zv9/danRer7)
• Human: Nov 4, 2011 (NCBI36/hg18)

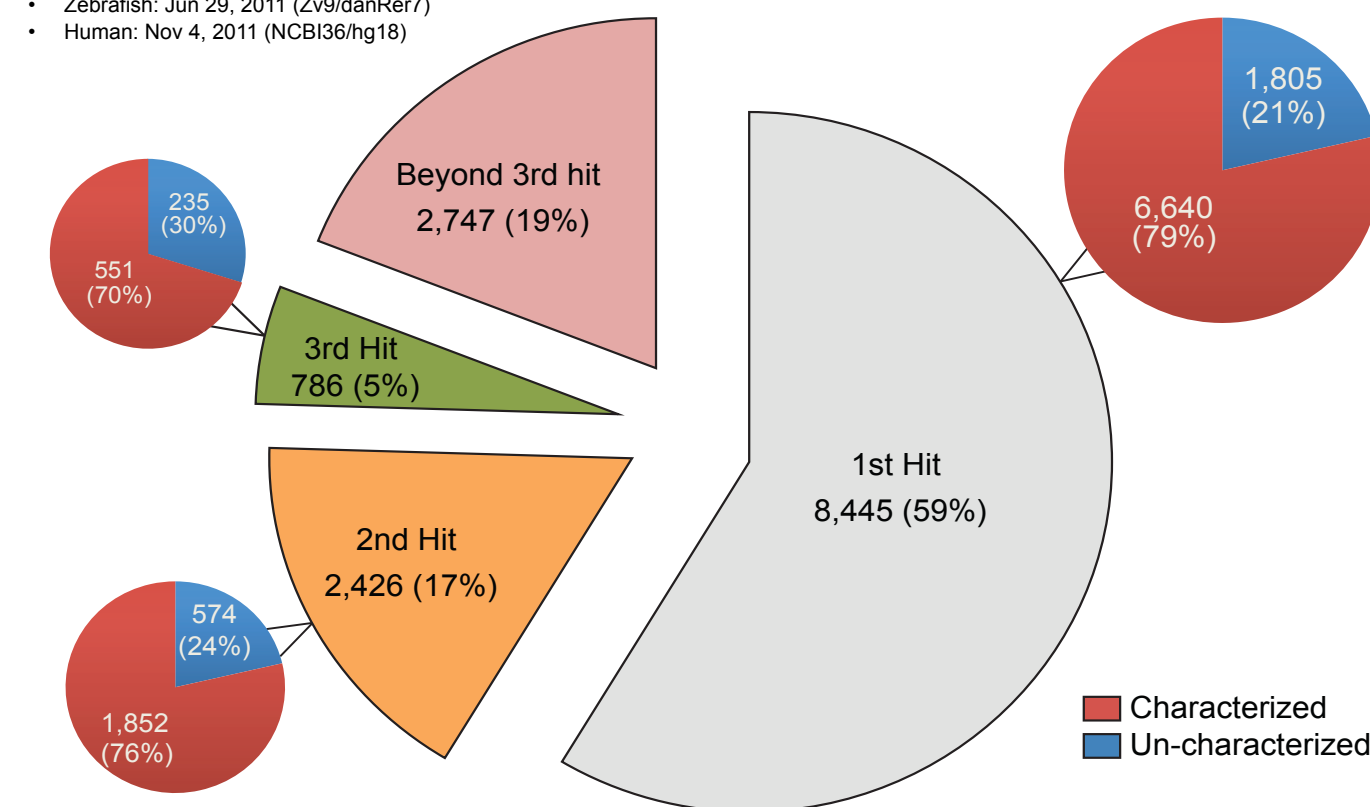


Figure 2. Results of the characterization of human orthologs. Within each orthology category, the number of characterized and un-characterized genes is reported.

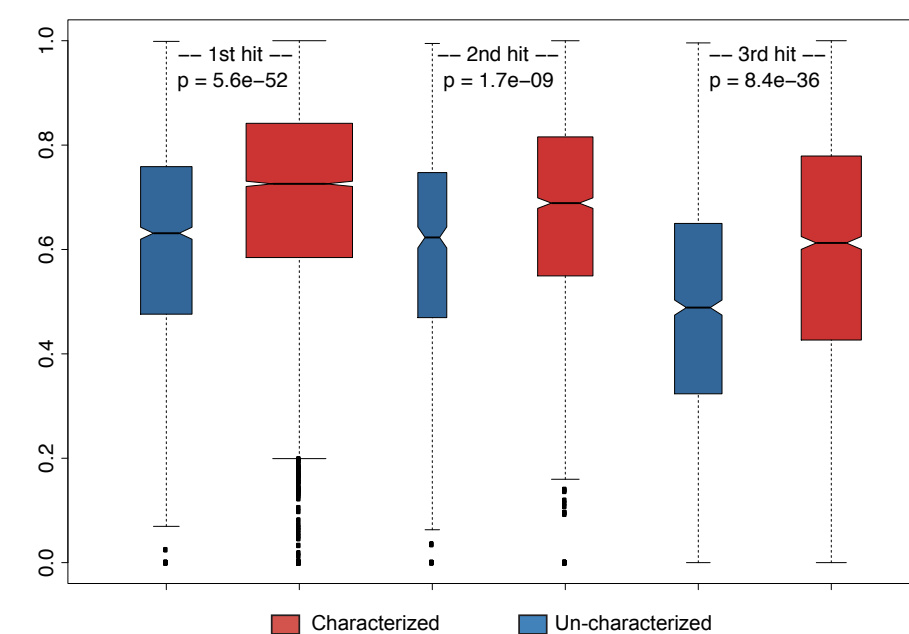
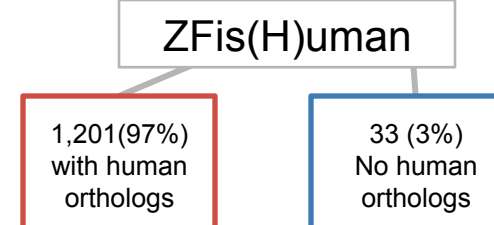


Figure 3. Conservation. Comparison of conservation scores measured by PhastCons¹⁻³ across 8 species, between characterized and un-characterized genes in each orthology category.

Table 4. Comparison of RefSeq and Ensembl. The vast majority of the common un-characterized genes between RefSeq and Ensembl, can be associated to a putative human ortholog using ZFis(H)uman.

		RefSeq	
		Characterized	Un-characterized
Ensembl	Characterized	11,047	1,778
	Un-characterized	262	1,234



Conclusions

In summary, we describe the implementation of a computational approach to identify human orthologs for the majority of zebrafish genes. This information should facilitate functional analysis of specific zebrafish genes and facilitate genomic analyses on larger datasets. We envision that the integration of this approach with the latest annotation sets along with a synteny based analysis, as applied by Barbazuk et al.⁴, can further improve our ability to identify orthologous relationships between zebrafish and human genes.

Results

Statistics

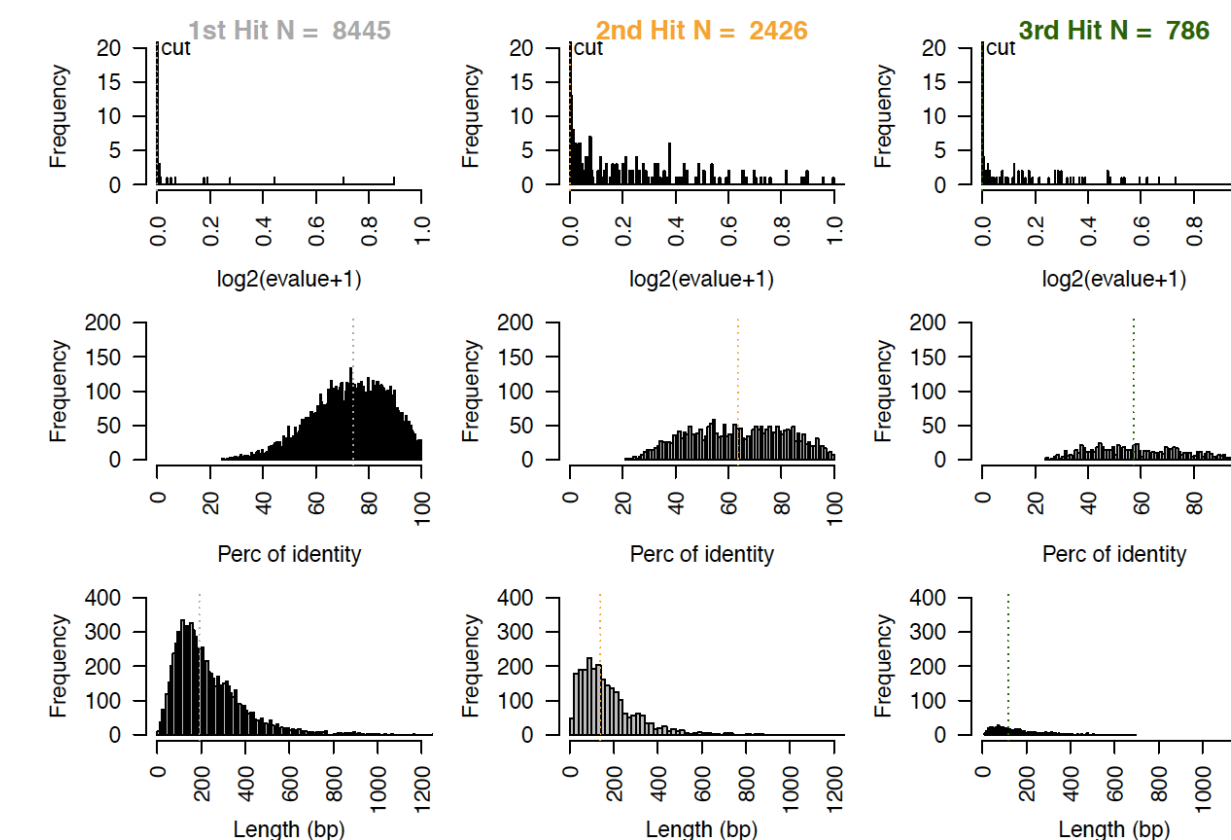


Figure 4. Histograms of the statistics for each orthology category. The median value is represented by a vertical line.

Top 100 differentially expressed genes

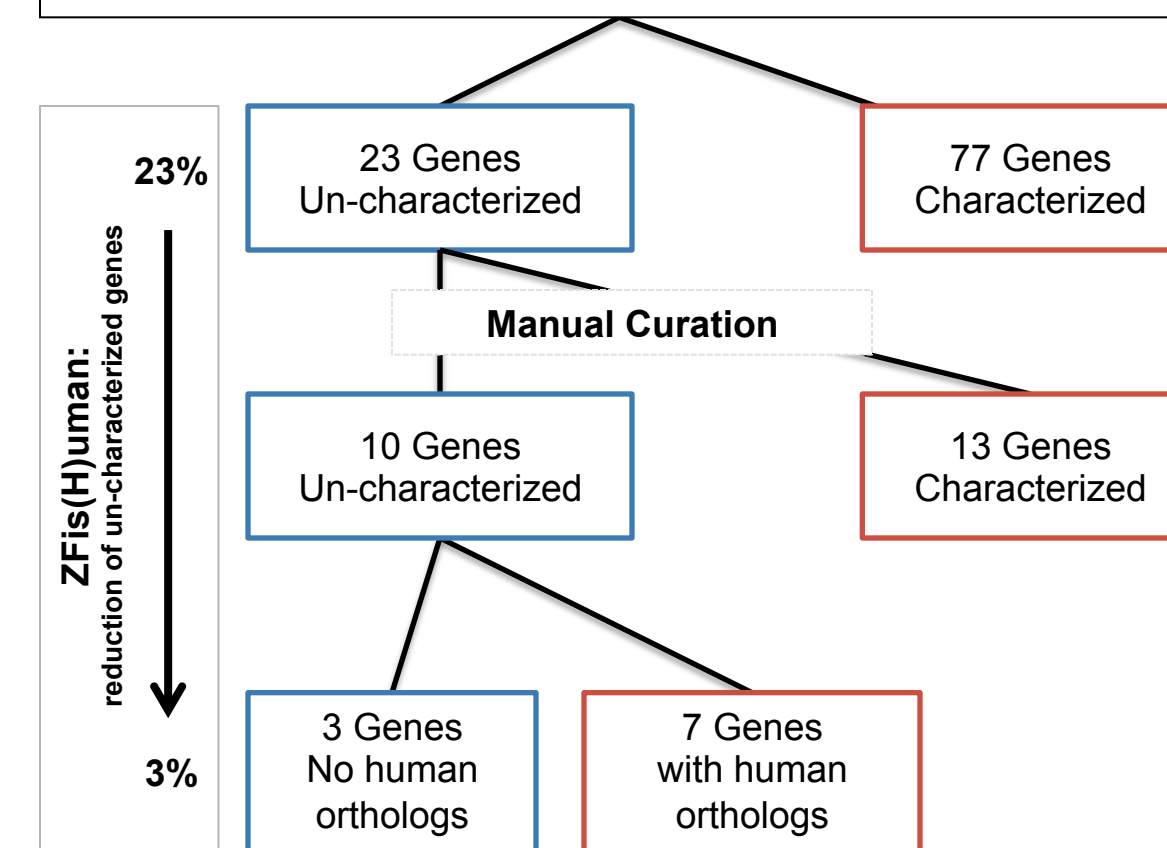


Figure 5. Practical impact. ZFis(H)uman identifies human orthologs for almost all the top 100 differentially expressed genes (between *tub* and *nacre*). Moreover, ZFis(H)uman identifies a human ortholog for 7 out of 10 genes that remain uncharacterized after manual curation.

References & Acknowledgements

To download the list of zebrafish-human orthologs, please visit:
<http://icb.med.cornell.edu/go/zebrafish>

- Yang Z. J Mol Evol 1994 Sep;39(3):306-14
- Felsenstein, J, and G A Churchill Mol Biol Evol 1996 Jan;13(1):93-104
- Siepel et al. Genome Res 2005 Aug;15(8):1034-50
- Barbazuk et al. Genome Res 2000. 10: 1351-1358

We would like to thank Dr. Yi Zhou for his useful suggestions.

