Accessing the Phenotype Gap: Enabling Systematic Investigation of Paralog Functional Complexity with CRISPR

Ben Ewen-Campen,[1] Stephanie E. Mohr,[1,2] Yanhui Hu,[1,2] and Norbert Perrimon[1,3,*]
[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]Drosophila RNAi Screening Center, Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]Howard Hughes Medical Institute, Boston, MA 02115, USA
*Correspondence: perrimon@receptor.med.harvard.edu
https://doi.org/10.1016/j.devcel.2017.09.020

Single-gene knockout experiments can fail to reveal function in the context of redundancy, which is frequently observed among duplicated genes (paralogs) with overlapping functions. We discuss the complexity associated with studying paralogs and outline how recent advances in CRISPR will help address the ''phenotype gap'' and impact biomedical research.

Paralogs, Redundancy, and the ''Phenotype Gap''

For nearly all organisms, genetic redundancy helps to ensure robustness in a variable world. However, there is one particular group of organisms for which genetic redundancy can be a major pain, too: namely, geneticists. Redundancy poses a major challenge for loss-of-function (LOF) studies, a powerful tool for uncovering gene function. When multiple genes function redundantly, knocking out any one individually does not produce a phenotype, and thus the functions of those genes remain invisible. The large number of genes without a detectable loss-of-function phenotype has been called the ''phenotype gap.''

A major source of functional redundancy is gene duplication, which results in the birth of a paralog. One might suppose that a newborn paralog would quickly be lost to relaxed selection unless it acquired a novel, advantageous function. And indeed, analyses of the fruit fly genome indicate that of the ~80 duplication events that occur every 1 million years, 96% have been lost to degradation (Rogers et al., 2009), a process termed ''nonfunctionalization.'' However, for the small fraction of new paralogs that do survive, three evolutionary trajectories become available: the acquisition of novel functions (neofunctionalization), the retention of varying degrees of overlapping function (subfunctionalization), or a combination of the two (Rogers et al., 2009).

Given the outsized role that duplicated genes play in generating evolutionary novelty, paralogs have intrigued evolutionary biologists for decades. In this Commentary, however, we focus on paralogs not as a substrate for evolution, but rather as a technical hurdle for functional genetic studies. Although attempts have been made to disrupt large numbers of paralogs in yeast, we argue that until recently it has not been technically feasible to systematically interrogate redundant paralogs in multicellular model organisms. As a result, many such genes likely remain uncharacterized, and for others we likely have an incomplete picture of their function.

For example, although *Drosophila* is arguably the most intensively studied multicellular genetic model system, an estimated 6,632 of the 13,919 genes in the genome (~48%) do not have annotated LOF phenotypes. This calculation is based on the phenotype annotations in FlyBase r6.16 and includes all protein-coding genes with no phenotype annotation, or annotated solely as ''viable'' or ''fertile.'' The phenotype data in FlyBase is culled from the literature and includes classical and modern mutagenesis screens, as well as more recent techniques such as RNAi screens. Of these, 2,303 (~35%) have paralogs. (We defined a list of 9,362 *Drosophila* paralogs using the DIOPT tool [http://www.flyrnai.org/diopt-dist v6.01] and further narrowed this list to include only those with a DIOPT score > 1 [7,152 genes], those that score as reciprocal best hits [6,173 genes], and those with <4 paralogs, for a final list of 5,463 paralogs.) We propose that the lack of detectable LOF phenotypes for a portion of these genes is due to paralog-based redundancy. Given the availability of highly effective CRISPR-based techniques for generating multiplexed knockouts simultaneously, as well as scalable new tools for overexpression experiments, we argue that it is now possible to systematically apply genetic approaches to the study of redundant paralogs in the genomes of multicellular organisms, which will help to reduce the size of the phenotype gap.

Redundancy Is Complex and Context Specific

Within every eukaryotic genome, there are large numbers of paralogs, and there is ample evidence that such paralogs contribute to redundancy. For example, large-scale studies in yeast show that when both members of a yeast paralog pair are knocked out, the phenotypic effects are significantly greater than additive, suggesting that functional redundancy is widespread (Dean et al., 2008). And importantly, there are many instances of paralogs retaining redundant function over hundreds of millions of years (Vavouri et al., 2008).

Despite these discernable trends, however, on a case-by-case basis paralogs often have complex and context-dependent relationships with one another. Genetic dissection of the Wnt pathway in *Drosophila*, for example, illustrates several ways in which redundancy can be complex and context specific. In this pathway, both the receptors (Frizzled [Fz] proteins) and the ligands (Wnt proteins)

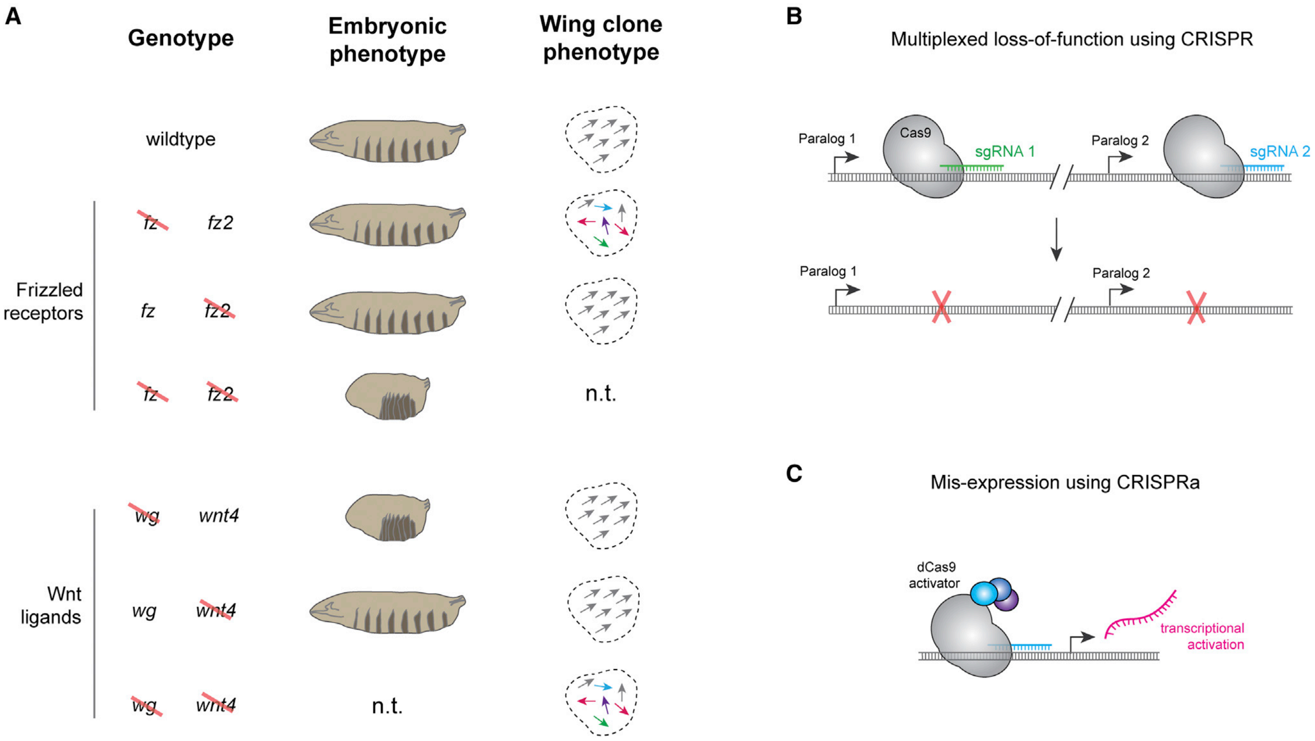

Figure 1. CRISPR-Based Tools for Studying Complex, Context-Dependent Redundancy

(A) Redundancy in the Wnt/Fz pathway. In the context of embryonic development, *fz* and *fz2* are fully redundant, and the characteristic Wnt pathway cuticle phenotype is only seen in the double mutant (loss of cuticle between denticle belts). However, *fz,* but not *fz2*, has a non-redundant role in PCP in the wing (gray arrows indicate properly oriented wing epithelial cells; colored arrows represent misoriented cells that have lost proper polarity). For Wnt ligands, whereas *wg* has a well-characterized LOF phenotype in embryos, its role in PCP in the wing is redundant with *wnt4*. n.t., not tested; refers to genotype-phenotype analyses that have not been published, to our knowledge.

(B) Multiplex CRISPR mutagenesis is a scalable technique for simultaneously disrupting multiple genes at once.

(C) CRISPRa allows for targeted overexpression of any gene of interest from the endogenous loci.

are members of paralog families that arose and diversified early in animal evolution, and they illustrate important nuances to our understanding of redundancy.

During embryonic development, Fz and Fz2 play seemingly identical roles in transducing signaling by the fly Wnt ligand Wingless (Wg) (Figure 1A) (Bejsovec, 2006). In fact, due to their functional redundancy, the roles of Fz and Fz2 as Wg receptors remained undiscovered for years after the other core components of the pathway had been identified. It was not until *fz fz2* double-mutant embryos were generated that their roles were definitively established (Figure 1A) (Bejsovec, 2006).

However, in other biological contexts, Fz and Fz2 perform entirely non-redundant functions. For example, Fz, but not Fz2, plays a major role in establishing planar cell polarity (PCP) (Figure 1A)—in fact, PCP was the namesake mutant phenotype for *frizzled*, owing to the aberrant bristle phenotypes in *fz* adult flies (Bejsovec, 2006). In addition, two other paralogs, Fz3 and Fz4, both bind to various Wg paralogs *in vitro* (Wu and Nusse, 2002), yet both genes have largely unknown biological functions. Furthermore, a distantly related *fz* paralog, *smoothened*, does not bind to Wg proteins at all and is instead a central component of Hedgehog signal transduction (Wu and Nusse, 2002). Thus, the *fz* family presents several examples of how paralogs can be redundant in one context but not in others.

Like their receptors, the seven *Drosophila* Wnt ligands also form a family of paralogs with complex overlapping functions and redundancies. For example, *wg* and *wnt4* are redundantly required for proper PCP signaling in the *Drosophila* wing; neither single knockout displays a PCP phenotype, whereas the double knockout does (Wu et al., 2013) (Figure 1A). This finding is particularly interesting because *wg*, and to a lesser extent *wnt4*, have both been extensively studied in other contexts. Thus, even in cases where a gene has well-characterized non-redundant functions, there may be hidden redundancy that is only detectable in specific cell types and/or at specific stages. In vertebrate genomes, the picture is likely to be far more complex, because both the Wnt and Fizzled paralog families have expanded greatly.

Complex, context-dependent interactions between or among paralogs are likely to be the rule, not the exception. In fact, theory predicts that maintaining some amount of overlapping function may account for the long-term evolutionary stability of paralogs (Vavouri et al., 2008). What's more, although paralogs are traditionally understood to provide robustness via functional redundancy, a recent study of 56 paralog pairs in yeast found that, surprisingly, the deletion of one paralog was equally likely to disrupt the function of its partner paralog as it was to be compensated by that partner (Diss et al., 2017). In those cases where paralogs have a mutually dependent relationship, this dependency often appears to be driven by physical interactions between paralogs (Diss et al., 2017). Altogether, we find compelling reasons to systematically study the functions of paralogs on a case-by-case basis by taking advantage of CRISPR techniques that make this newly feasible (discussed below).

Paralog Biology Has Biomedical Implications

Over 80% of human disease-linked genes have paralogs, spanning a wide range of duplication dates (Dickerson and Robertson, 2012). Thus, in addition to contributing to functional annotation of individual genes, gaining a more thorough understanding of paralog biology will likely have important implications for human disease research and treatment. A number of recent genetic screens in cancer cell lines have identified synthetic lethal interactions between paralogous oncogenes, including several paralogous components of the SWI/SNF chromatin remodeling complex members (O'Neil et al., 2017). In these cases, a paralog that is dispensable in wild-type cells becomes essential in cancer cells that are mutant for its partner paralog, thus providing a potential new therapeutic target.

Interestingly, there are also now multiple examples of cancer-causing chromosomal deletions that encompass "passenger genes": genes that are incidentally deleted alongside a tumor suppressor, that are members of paralogous gene families, and whose incidental deletion creates a cancer-specific vulnerability. For example, pancreatic ductal adenocarcinoma (PDAC) cancer cells often contain homozygous deletions of the causal tumor suppressor, *SMAD4*, as well as the neighboring gene mitochondrial enzyme *malic enzyme 2* (*ME2*). Because these PDAC cells also lack *ME2*, they become highly sensitized to loss of its paralog, *ME3*, thus providing a novel therapeutic target (Dey et al., 2017).

Paralogy is likely to inform drug design in multiple additional ways. Whereas it is typically desirable to minimize drug cross-reactivity, there is now growing recognition that, in some cases, it may be advantageous to promiscuously target multiple paralogs to more broadly inactivate biological activity and prevent compensation. The RAF kinases provide one example. There are three paralogous RAF kinases in humans (ARAF, BRAF, and RAF1, also known as CRAF), of which BRAF is the most frequently mutated in cancers. In the context of malignant melanoma, two drugs that specifically target a common oncogenic BRAF allele (BRAF^{V600E}) often lead to resistance within a matter of months via multiple mechanisms (Girotti et al., 2015). In contrast, two compounds were recently developed that target both BRAF^{V600E} and RAF1, as well as receptor tyrosine kinase/SRC-family kinase (SFK) signaling, and these compounds appear to be effective and to avoid resistance (Girotti et al., 2015).

Another example involves epidermal growth factor (EGFR) signaling. Hyperactivation of this signaling pathway is a common feature of many types of cancer, and drugs targeting the EGFR protein were among the earliest anti-cancer therapies. However, the EGFR protein has three paralogs that are also involved in this pathway (HER-2, HER-3, and HER-4), and one of the most frequent mechanisms of resistance to EGFR-specific inhibitors involves amplification of HER-2 (Milik et al., 2017). Thus, new compounds are being developed that target both EGFR and HER-2 in an attempt to overcome this acquired resistance (Milik et al., 2017).

As these examples illustrate, for well-studied examples, paralog biology has already informed disease research and drug design. Expanding our understanding of the role of redundancy to under-studied cases therefore holds promise.

New Techniques for Studying Redundant Genes

Characterizing redundant genes benefits from two complementary genetic approaches: multi-gene knockout LOF studies and single- or multi-gene overexpression studies. Double-LOF experiments are critical for demonstrating that two redundant genes are necessary in a given process, as illustrated above by the example of Wnt/Fz paralogs. Overexpression experiments can provide insights into a gene's function and can be particularly informative for genes that lack a LOF phenotype. For example, prior to the creation of *fz fz2* double-mutant embryos, overexpression studies *in vitro* had shown that Frizzled proteins are capable of transducing the Wg signal (Nusse et al., 1997). Similarly, the initial indication that *wg* and *wnt4* both influence PCP in the fly wing came from overexpression experiments, which were followed by the more laborious creation of double-mutant clones (Wu et al., 2013).

While neither double-LOF nor overexpression is a new technique, CRISPR has revolutionized the ease, speed, and scalability of both. Until recently, the creation of double mutants has been time consuming and arduous, even in model organisms such as *Drosophila*. Thus, large-scale screens for genetic interactions have traditionally been conducted in a single genetic background, looking for enhancers or suppressors of a specific phenotype, rather than looking at many different one-plus-one knockouts. Double-RNAi techniques were a dramatic improvement in terms of scalability, but they also suffer from the compounding effects of incomplete knockdown and off-target effects frequently observed with RNAi.

CRISPR makes it possible to disrupt any two (or more) genes of interest simultaneously in specific cells using approaches that can be applied at large scale (Figure 1B). For example, Port and Bullock simultaneously knocked out four genes *in vivo* in fruit flies using a single, easily cloned transgene encoding four short guide RNAs (sgRNAs) in tandem (Port and Bullock, 2016). Thus, researchers are now able to screen not only for genetic interactions in a single mutant background but also in high throughput for genetic interactions among many pairs of genes. The massively parallel double-LOF approach has been demonstrated in mammalian cell culture using both CRISPR and CRISPRi (O'Neil et al., 2017; Du et al., 2017). In these studies, a large library of "dual-guide" constructs, each encoding two sgRNAs targeting separate genes, are introduced to cells expressing Cas9, and the cell population is then screened in a pooled format. To date, a number of screens using dual-sgRNA library, representing tens of thousands of combinations of drug targets, have been used to identify strong and reproducible genetic interactions, both synergistic and buffering, that correspond to new and effective drug combinations (O'Neil et al., 2017).

Similarly, there are now a number of powerful CRISPR-based gain-of-function (GOF) techniques based on fusing transcriptional activation domains to catalytically dead Cas9 (CRISPRa approaches) that allow for targeted misexpression from endogenous loci (Chavez et al., 2016) (Figure 1C). Importantly, although these techniques have been primarily used in cell culture thus far, they are rapidly being adapted for *in vivo* double-KO and GOF studies, where they can be combined with inducible expression systems for spatial and temporal control and can be scaled up to generate large-scale resources (Ewen-Campen et al., 2017). These newly available techniques should contribute greatly to our ability to systematically interrogate the functions of redundant paralogs both *in vitro* and *in vivo*.

Conclusions

Paralog-based redundancy is widespread and presents a fundamental challenge to traditional LOF analyses. And importantly, although there has been much theoretical work on the evolution and diversification of paralogs, there is no reason to believe that the functional relationship between two given paralogs will be predictable from general trends. Rather, scalable tools are needed to screen paralogs and subsequently characterize them in detail on a case-by-case basis with spatial and temporal control. The availability of powerful new CRISPR-based tools opens the door to systematic study of redundant paralogs via double-KO and GOF studies. Given what has been uncovered already regarding paralog functions in signal transduction and cancer biology, the systematic study of paralogs is likely to provide important new insights into basic and applied biology.

REFERENCES

Bejsovec, A. (2006). Oncogene *25*, 7442–7449.

Chavez, A., Tuttle, M., Pruitt, B.W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S.J., Cecchi, R.J., Kowal, E.J.K., Buchthal, J., et al. (2016). Nat. Methods *13*, 563–567.

Dean, E.J., Davis, J.C., Davis, R.W., and Petrov, D.A. (2008). PLoS Genet. *4*, e1000113.

Dey, P., Baddour, J., Muller, F., Wu, C.C., Wang, H., Liao, W.-T., Lan, Z., Chen, A., Gutschner, T., Kang, Y., et al. (2017). Nature *542*, 119–123.

Dickerson, J.E., and Robertson, D.L. (2012). Mol. Biol. Evol. *29*, 61–69.

Diss, G., Gagnon-Arsenault, I., Dion-Coté, A.-M., Vignaud, H., Ascencio, D.I., Berger, C.M., and Landry, C.R. (2017). Science *355*, 630–634.

Du, D., Roguev, A., Gordon, D.E., Chen, M., Chen, S.H., Shales, M., Shen, J.P., Ideker, T., Mali, P., Qi, L.S., and Krogan, N.J. (2017). Nat. Methods *14*, 577–580.

Ewen-Campen, B., Yang-Zhou, D., Fernandes, V.R., González, D.P., Liu, L.-P., Tao, R., Ren, X., Sun, J., Hu, Y., Zirin, J., et al. (2017). Proc. Natl. Acad. Sci. USA *114*, 9409–9414.

Girotti, M.R., Lopes, F., Preece, N., Niculescu-Duvaz, D., Zambon, A., Davies, L., Whittaker, S., Saturno, G., Viros, A., Pedersen, M., et al. (2015). Cancer Cell *27*, 85–96.

Milik, S.N., Lasheen, D.S., Serya, R.A.T., and Abouzid, K.A.M. (2017). Eur. J. Med. Chem., S0223-5234(17)30545-7.

Nusse, R., Samos, C.H., Brink, M., Willert, K., Cadigan, K.M., Wodarz, A., Fish, M., and Rulifson, E. (1997). Cold Spring Harb. Symp. Quant. Biol. *62*, 185–190.

O'Neil, N.J., Bailey, M.L., and Hieter, P. (2017). Nat. Rev. Genet. *18*, 613–623.

Port, F., and Bullock, S.L. (2016). Nat. Methods *13*, 852–854.

Rogers, R.L., Bedford, T., and Hartl, D.L. (2009). Genetics *181*, 313–322.

Vavouri, T., Semple, J.I., and Lehner, B. (2008). Trends Genet. *24*, 485–488.

Wu, C.-H., and Nusse, R. (2002). J. Biol. Chem. *277*, 41762–41769.

Wu, J., Roman, A.-C., Carvajal-Gonzalez, J.M., and Mlodzik, M. (2013). Nat. Cell Biol. *15*, 1045–1055.