UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXVI Ciclo

Understanding the mechanisms of transcriptional robustness during embryonic development.

Enrico Cannavo'

PhD Thesis

Scientific tutors: Eileen Furlong and Roberto Mantovani

Academic year: 2010-2013

SSD: BIO/18

Thesis performed at European Molecular Biology Laboratory (EMBL)

Index

Index 1

Chapter 1

Chapter 2

Chapter 3

Chapter 1

Abstract

Embryos develop into stereotypically patterned organisms that are largely robust to differences in segregating mutations and environmental conditions. Such robustness is in part conferred through elements that perform similar, and thus partially redundant, tasks. Partially redundant enhancers, for example, confer precision and robustness to gene expression during development, as shown at individual, well-studied loci. However, the extent to which enhancer redundancy exists and can thereby have a major impact on developmental robustness remains unknown. Here, we systematically address this question, identifying over 2,000 predicted pairs of partially redundant enhancer elements (PREEs) during *Drosophila* mesoderm development. The activity of 28 specific elements, distributed throughout 7 loci, was compared in transgenic embryos, while natural sequence variation – i.e, structural variation – among ‘individuals’ was used to assess their potential redundancy. Our results reveal three clear properties of enhancer redundancy within developmental systems. First, enhancer redundancy is much more pervasive than previously anticipated, with 70% of loci examined having two or more PREEs. Second, over 50% of tested loci do not follow the simple situation of having only two redundant elements as generally assumed – often there are three (*rols*), four (*CadN* and *ade5*) or even five (*Traf1*) enhancers with overlapping spatio-temporal activity, where at least one of which can be deleted without obvious phenotypic effects. Third, this level of potential robustness in transcriptional regulation is not reserved for the key developmental regulators, or selector genes, but rather many genes regardless of their function have extensive levels of *cis*-regulatory redundancy.

Introduction

Living organisms are extremely complex and can respond in different ways to internal perturbations, such as genetic mutations, and external variations, like environmental perturbations. Phenotypic responses to different environments may be consistent among genotypes, which in a developmental context this robustness is called canalization. Alternatively, genotypes may have different plasticities, indicating that interactions between genotype and phenotype are not always one to one: in some cases under different perturbations, a single genotype can give rise to different phenotypes (phenotypic plasticity) while in other cases different genotypes bring about a similar constant phenotype despite genetic and environmental variations (canalization). Those phenomena have been extensively described. Here I will discuss a few examples of phenotypic plasticity and then more extensively about robustness during embryonic development and how evolution uses one or the other depending on the context. I will particularly focus on the mechanisms that are responsible to ensure robustness of an organism's development and how transcriptional redundancy in different species plays an important role in regulating robust and precise gene expression during embryonic development.

Phenotypic plasticity

The relationship between genotype-and-phenotype is not one::one. Rather, a single genotype can produce different phenotypes in response to environmental stimuli (Fig. 1). This phenomenon is called phenotypic plasticity and refers to situation in which the same set of genes can yield different phenotypic outcomes when exposed to distinct environmental conditions. It includes rapid reversible changes in behavior, physiology and morphology. Several cases of animals immediately changing their behaviors or morphology in a new environment context have been described. An example of behavioral plasticity has been observed between several species of tits (*Parus* spp.) that coexist in the same habitat. In particular, the competition for foraging sites between Coal Tits (*Parus ater*) and Willow Tits (*P. montanus*) in Swedish coniferous forests has been investigated (Alatalo and Moreno 1987). Coal tits are generally smaller (9.5 gr) then the willow tits (11.5 gr). In a mixed environment where both species are present, the willow tits forage on the inner tree parts while the coal tits forage on the outer tree parts, because of the social dominance that the larger species have over the smaller ones (ALATALO et al. 1985). Alatalo RV and Moreno J (Alatalo and Moreno 1987) demonstrated that in laboratory conditions that mimic an absence of willow tits, coal tits now tended to prey on food items in the inner tree parts. When the environment was perturbed by the introduction of willow tits, the coal tits underwent a behavioral change that pushed them to prey at the outer tree parts (Alatalo and Moreno 1987). These observations were confirmed by experiments in the field. However, behavioral plasticity is a really broad concept that could include most human behaviors and all learning processes of animals in general.

Here I want to focus on examples of phenotypic plasticity that produces irreversible morphological changes of an organism and that involve developmental reactions. The *Pseudocrenilabrus multicolor*, known as the Egyptian Mouthbrooder, is a small mouth-brooding haplochromine cichlid fish. It lives in a wide range of habitats that differ widely in stability and oxygen availability. It has been found that environmental perturbations, i.e. different conditions of oxygen availability, cause morphologically changes in the gills and some nearby muscle structures of the fish (Chapman et al. 2000).

It is as well possible that morphological and behavioral differences can be directly connected. Two populations of hummingbirds (*Amazilia tobaci*) were described in two different South American islands, Tobago and Trinidad, with few environmental differences (FEINSINGER and SWARM 1982) aside from the number of competitors, 4 and 6 respectively on Tobago and Trinidad. In line with this, the only detectable morphological difference was related to the wing size that correlated with the ability to defend rich food resources from several competitors (FEINSINGER and SWARM 1982). This ability also associated with the feeding habits of the two populations, especially during period of food shortage (FEINSINGER and SWARM 1982).

A clearer case of morphological and behavioral differences is represented by two related species of stickleback (*Gasteosteus* spp). Commonly referred to “benthic” and “limnetic” because of their diet habits, they are extremely close species that derived from the marine stickleback ancestor (Schluter and McPhail 1992). The trophic morphology explains most of the foraging behavior and is largely hereditable (McPhail

1992). Raising benthic fish on limnetic diet from early stages, it caused the displacement of benthic morphology towards the limnetic one and vice versa (Day et al. 1994). Therefore, both species exhibit adaptative morphological plasticity that mostly depends on the food availability and variability (Day et al. 1994). Moreover this study suggested that differences between two closely related species are the result of both genetic and environmental influences (Day et al. 1994).

Taken together these studies offer evidence for how organisms respond to a phenotypic change following an environmental alteration. Such plasticity has a clear selective value: In the face of fluctuating environments, phenotypic plasticity allows individual genotypes to match to multiple phenotypic optima. There is thus general agreement that phenotypic plasticity, as a trait, can be acted upon by selection (Pigliucci et al. 2006). However, it has also been argued that phenotypic plasticity can be a *driver* of adaptation. This means that adaptation would be initiated by phenotypic plasticity and not by selection on genetic variations. This has been a very controversial topic so far (De Jong 1995; Via et al. 1995; Pigliucci et al. 2006). There are several models that attempt to explain a potential role of phenotypic plasticity as an evolutionary driving force (de Jong 2005). Trying to include all previously published models, de Jong (de Jong 2005) concluded that "plasticity is itself not an evolutionary mechanism and does not promote different evolutionary solutions" since in the model phenotypic plasticity is predicted as a consequence of previous selection. However, Pigliucci *et al*. (Pigliucci et al. 2006) claimed that plasticity is a mechanism "in the sense of approximate cause" of changes and the fundamental cause of adaptation during evolution is the natural selection. Thus, it is likely that phenotypic plasticity helps adaptation to a new environment if it points towards a new optimum (Futuyma 2009).

On the other end, Waddington showed how environmentally induced phenotypic characters fixed in the population after a process of selection going on for several generations (Waddington 1953). This suggests that phenotypic plasticity could promote the expression of new phenotypes that are then selected to improve the performance of the population resulting in the phenomenon called genetic assimilation (Pigliucci et al. 2006; Waddington 1953). Waddington (Waddington 1953) performed an experiment that showed how, in lab conditions, he could induce phenotypic plasticity and the acquisition of a new trait in *D. melanogaster*, which was maintained in a population even in the absence of environmental disturbance. In detail, wild-type *D. melanogaster* pupae were heat shocked at 40°C (Waddington 1953). Already after 5 generation, he could observe two selected populations of flies: one where no new phenotypes were detected ("normal") and another where the development of the wing crossvein was affected ("crossveinless") (Waddington 1953). Those phenotypes were maintained also when flies were raised under normal conditions (Waddington 1953). Waddington proposed that if the same experiment was repeated with another population of flies with different genetic background, the same phenotypic change could be observed. Thus, he found that selection could fix a phenotypic change that was initially induced by environmental persistent perturbations, transforming environmentally-induced phenotypic variation into genetically encoded differences between populations.

For many years, though, this phenomenon has been criticized, especially from Williams' (Williams 1966) who rejected the role of genetic assimilation. He claimed that if certain extreme situations become recurrent or permanent, any phenotypic change would be eliminated rather than assimilated by natural selection (Williams 1966; Eshel and Matessi 1998).

However the phenotypic plasticity remains a very complex and slow process achieved by natural selection. The phenotypic evolution was proposed to result from four steps (West-Eberhard 2003): environmental variations cause the appearance of a new phenotype (*trait origin*) that determines the rearrangement of the different aspect of the phenotype (*phenotypyic accommodation*) to the new trait and followed by an *initial spread* of the new phenotype. In this way a small subpopulation of individuals express the new trait that will be at the end fixed by selection (*genetic accomodation*). Thus, once a trait has become plastic, it could then be the target of genetic assimilation, assuming a persistent environmental destabilization (Eshel and Matessi 1998).

The Caribbean *Anolis* lizards may represent a possible example of phenotypic plasticity followed by genetic assimilation. The length of *Anolis* lizards' limbs shows great variations between individuals of the same or different species. This variation is associated with structural differences in the habitat. Taxa found in habitats with broad surfaces tend to have longer hind limbs than the relative species that uses narrow surfaces which has short limbs (Losos et al. 2000). To test if the evolution of the limbs was the result of phenotypic plasticity and genetic assimilation or fixation amongst the population, Losos *et al*. (Losos et al. 2000) grew *Anolis sagrei* in two different environments, with narrow or broad perches. They observed that the limbs grew to significantly different lengths in the two environments following the same trend that had been seen before (Losos et al. 2000). The authors claimed that this was the first demonstration of phenotypic plasticity in vertebrates and, together with genetic assimilation and natural selection, it could have played an important role in the evolutionary radiation of *Anolis* lizards in the Caribbean.

The flowering plant *Achillea luminosa* presents another potential example where morphological differences in their leaves was explained by genetic divergence and phenotypic plasticity. The effects of growing *Achillea* plants collected from an intermediate and a high altitude site under contrasting temperature was investigated (Gurevitch 1992). The size and compactness of the leaves changed depending on the altitude and the temperature (Gurevitch 1992). Differences between the two populations were maintained if they were grown under the same temperature and altitude conditions (Gurevitch 1992). This is a good example of how adaptive phenotypic plasticity is responsible for a change in a particular trait, which is maintained in the population despite the absence of environmental disturbances.

I described above the case of the two species of stickelbacks that respond with morphological and behavioral changes to environmental variations. In this case, since phenotypic plasticity was still detected in both species (Day et al. 1994), it is possible that genetic assimilation is still in progress.

Developmental robustness

Natural selection is slowly and constantly acting to select those organisms that have a better fitness and, by extension, can better adapt to a new environment. This can be reached in two different ways: (1) as described previously, an organism is able to quickly modify its phenotype when environmental perturbations occur; (2) a biological system produces an invariant output or phenotype despite internal or environmental changes (Fig. 1). Waddington (Waddington 1942) described the latter phenomenon and coined the term canalization. In his terms, reactions are canalized when “they are adjusted [by natural selection] so as to bring about one definite end-result regardless of minor variations in conditions during the course of the reaction” (Waddington 1942). This means that when a reaction is well canalized it is also robust: most genetic and environmental variation have little effect on gross phenotypes so that a population remains phenotypically uniform but genetically heterogeneous.

The nature and the basis of canalization have been well described by several studies. A classical well-studied case is the number of bristle in the scutellum of *D. melanogaster*. Their development depends on the *scute* (*sc*) gene (RENDEL and SHELDON 1960). In wild-type flies the number of bristles is stereotypical and fixed at 4 and the variance between individuals is very low (RENDEL and SHELDON 1960). Mutant and wild type flies underwent selection for higher number of bristles and after several generations this number was brought back to 4 in mutant flies while it increased significantly for wild type flies. In these conditions the wild-type flies showed a higher variability than mutants, indicating that canalization was acting to maintain the number of bristles to 4 and prevent large variation. Deviation from the normal number is observed only when the system is forced to move away from the median (RENDEL 1959).

In another study, the role of canalization on photoreceptor determination in *D. melanogaster* eye was investigated. The *Drosophila* compound eye is a set of approximately 800 ommatidia consisting of 8 photoreceptor cells (R1–R8). A tyrosine kinase receptor is important in the initiation of particular photoreceptor cell (R7) determination (Polaczyk et al. 1998). This process is constant in wild type individuals while it became variable in different gain-of-function mutants of the receptor (Polaczyk et al. 1998). The severity of the observed phenotype depended on the genetic background, suggesting that different alleles in genes that were part of the same signal transduction cascade could buffer the effect of mutations in the tyrosine kinase receptor (Polaczyk et al. 1998).

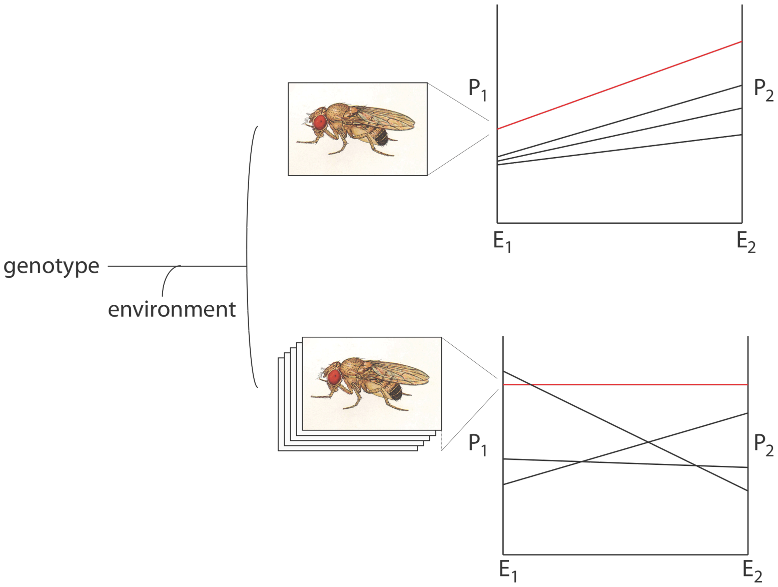


Fig. 1 Phenotypic plasticity and canalization. A single genotype under different environmental condition could produce different phenotypes (phenotypic plasticity, top) that in certain condition could be selected (red line) because close to the optimal. In other cases, though, a several different genotype could produce the same constant phenotype in different environmental conditions (canalization, bottom) that is favored by selection.

In quantitative genetics, canalization has been defined as any genetic change that results in a state of reduced variability of a trait (Wagner and Altenberg 1996). Disturbances that have an effect on the phenotype can be part of two categories: the external ones are those perturbations caused by environmental variations (environmental canalization) while the internal ones are changes at the genotype level, i.e. mutations (genetic canalization) (Wagner et al. 1997). While environmental changes are often temporary, mutations in genes are more permanent and can be transmitted to successive generations. Increases in phenotypic variance that result from extreme environmental conditions might be sufficiently deleterious to induce selection for those traits that do not have significant variations in most common environments (Eshel and Matessi 1998).

The fluctuating asymmetry of Australian sheep blowflies (*Lucilia cuprina*) is a classical example of robustness to environmental changes (Clarke and McKenzie 1987). Insecticides have been used to reduce the population of sheep blowflies, but resistance strains have been selected more quickly than expected. Initially those strains had lower fitness in comparison to the wild type and showed random differences between left and right body symmetry, an indication of developmental distress (Clarke and McKenzie 1987). After many generations and thereby accumulation of other mutations, the flies maintained the resistant phenotype and decreased the level of asymmetry to the normal level of insecticide-susceptible strain, suggesting that canalization acts on the resistant trait (Clarke and McKenzie 1987). This example shows how a population will move towards the least variable condition and how selection acts against individuals with non-optimal phenotypes so that canalization may evolve under the major force of stabilizing selection (Wagner et al. 1997).

However, it would be advantageous for systems if canalization was broken down when the environmental conditions become too extreme and in that case other phenomena can act on the selection of the organism (i.e. phenotypic plasticity, natural selection, genetic assimilation) (Eshel and Matessi 1998).

As already mentioned, environment is not the only perturbations against which canalization buffers. It can also operate at the genotype level, buffering developmental pathways against tendency of new alleles to make non-optimal phenotypes (Wagner et al. 1997). The genetic canalization, though, seems to be more difficult to picture because it is influenced by gene frequency and by their epistatic interactions (Wagner et al. 1997). Nevertheless there are well-studied cases that empirically described the nature of genetic canalization. An example is represented by the effect of the *Tabby* mutation on the whisker number in mice (DUN and FRASER 1958). Using the same experimental design that Rendel JM. and Sheldon BL. (RENDEL and SHELDON 1960) applied, the study focused on the number of secondary vibrissae in mice. A very stable number of 19 vibrissae in a survey of 3000 mice was observed. Heterozygous and homozygous mutations for the gene *Tabby*, however, showed an increase variance in vibrissae number relative to wild-type mice. Moreover, it was shown that those variations were hereditable and accumulation of genetic variation on selected mutants was sufficient to change the invariable wild type phenotype. These data lead the authors to conclude that there are genetic variation for vibrissae number in the background of the uniform genotype of wild type mice, revealed only in the mutant phenotype. This increase in variance was interpreted as implying that in wild type mice this trait is genetically canalized.

There are many other described cases in the literature that prove the existence of genetic canalization. I already described the case of *scute* in *D. malanogaster* (RENDEL and SHELDON 1960). It was clearly shown that canalization in this context is under genetic control and in a later paper the same authors were able to successfully shift the canalization from 4 to 2 bristles (RENDEL et al. 1965). They selected a line that had low variance in the number of bristles per generation and at different environmental temperature (RENDEL et al. 1965).

Several other studies aimed to investigate the role of genetic canalization in *D. melanogaster*, as reviewed by Scharloo W. (Scharloo 1991).

Genetic robustness

As defined earlier, genetic canalization or robustness is the ability to buffer the presence of genetic mutations from modifying or altering a particular trait. The evolutionary origin of genetic robustness is still not clear. Three different classes have been distinguished (VISSER et al. 2003): here I describe the (1) adaptative and (2) congruent robustness.

(1) Adaptative robustness refers to a situation in which robustness evolves because it increases the fitness of the genotype. Indeed, a well-adapted genotype might be negatively affected by any mutations and a mechanism that buffers against them should be favored by natural selection in a constant environment. Genetic canalization belongs to this class. The most important factors influencing the evolution of robustness is the mutation rate. Since it is more likely that mutations have detrimental effects on the organism (Eyre-Walker and Keightley 2007), a buffering mechanism is selectively advantageous. However the evolution of adaptative robustness also depends on other parameters that are linked to the mode of reproduction and to the population size (VISSER et al. 2003). Indeed, because the selective advantage of

robustness is in the same order of magnitude as sequence variation (VISSER et al. 2003), the population size should be big enough: this makes evolution of robustness less likely in small populations. Nevertheless, there have been studies claiming that evolution of robustness is possible thanks to the intervention of other mechanisms, such as redundancy and anti-redundancy (Krakauer and Plotkin 2002). The evolution of adaptative robustness depends also on the strength of stabilizing selection. This has two effects on adaptative robustness: on one hand it increases the fitness of the organism because of the strong selection on the robust trait but on the other hand it decreases the amount of genetic variation, lowering the onset of new phenotypes (Wagner et al. 1997). Therefore, the two effects neutralize each other so that trait under strong stabilizing selection in a constant environment may not evolve adaptative robustness.

(2) When the evolution of genetic robustness is a side effect of the evolution of environmental robustness, it is said to be congruent. The case of Hsp90 in *D. malanogaster* is an example (Rutherford and Lindquist 1998). Mutants affecting the function of this protein, which is important for the folding maturation of several proteins, show developmental abnormalities. In particular, crossing several wild type strains with Hsp90 heterozygous mutants caused the onset of different morphological abnormalities in the F_1 population (Rutherford and Lindquist 1998). This happened because cryptic genetic variations were expressed to a greater extent especially when flies were grown at extreme environmental temperatures. Thus, Hsp90 was observed to canalize the phenotype from temperature-dependent morphogenetic variants and to ensure robustness against mutational perturbations. Therefore, environmental robustness was first selected and with it, the role of Hsp90 in genetic robustness.

Congruent robustness can also increase the stability of regulatory systems through cooperativety. The kinetic representation of cooperative events is a sigmoid curve that has been proved to confer robustness to several perturbations (Masel and Siegal 2009).

Robustness and evolvability

By definition, genetic robustness reduces the impact of genetic variation on organismal phenotypes. This seems the opposite of what has been defined for evolvability: the capacity of a population to produce any kind of hereditable phenotypic variation that are not unconditionally deleterious (Masel and Trotter 2010). The two phenomena have been recently connected resolving what has been a long standing paradox . To understand that, it is easier to think about a genotype space, a representation of all possible existing genotypes in which two neighboring points differ for one single mutation. When all genotype sets of the space are connected to each other, they form a neutral network (Schuster et al. 1994). In this scenario, point mutations and structural variations are important since they can alter the neutral network modifying one genotype into one of the neighbors. Weakly deleterious mutations are more abundant than neutral mutations in most macromolecules (Eyre-Walker and Keightley 2007). If the rate of deleterious mutations is high and the phenotype is not conserved, selection will remove them from the neutral network. In this way the population will have a low diversity because of the extinction of many singular genotypes. Genetic variations, though, are followed by other buffering mechanisms that allow the persistency of the phenotype. Thus, robustness has a key role in ensuring the maintainment of the constant advantageous phenotype and at the same time allows the accumulation of mutations between two genotypes so that they will have more access to many more novel phenotypes when new perturbation (Wagner et al. 2012).

In the work of Wagner A. (WAGNER 2008), analyzing the secondary structure of the RNA molecules, it is concluded that there is a synergism between evolvability and robustness. He described, indeed, how high robust phenotypes have high evolvability. The larger is the neutral network, the greater is the robustness of the phenotype and in turn the higher is the evolvability. This indicates that under selection, one phenotype may be adopted by different genotypes that are close in the neutral network. When an organism is robust, cryptic mutations can accumulate without showing any phenotypic repercussion and at the same time those hidden mutations make the population spread out over a larger region of genomic space, increasing evolvability (WAGNER 2008). In conclusion, this work has demonstrated how robustness accelerates the evolvability, thanks to the accumulation of hidden mutations that are buffered by several strategies aiming a low variance phenotype despite several disturbances.

Mechanisms of robustness

Network motifs and robustness

Biological systems employ any different mechanisms to ensure robustness. Despite extensive mutational and environmental variation, nature has developed strategies that allow reactions to be carried till their "definite end-result". The presence of systems level control is one important mechanism associated with the acquisition of robustness. This typically consists of network motifs, defined as a small set of recurring patterns that regulate several biological processes from transcriptional regulation to chemotaxis (Alon 2007; Alon et al. 1999). They are particularly important to allow networks of interacting elements to change quickly in response to stimuli. In a simple regulation case, the autoregulatory mechanisms have been shown to increase the robustness of the network (Becskei and Serrano 2000).

Two motifs in particular can confer network robustness: negative and positive feedback.

In a negative feedback simple network, the system inhibits itself, i.e. a transcription factor (TF) represses the transcription of its own gene (for review Smale and Kadonaga 2003; Alon 2007). In a transcription network, the negative feedback of the TF on its own activity speeds up the kinetics of the response of the TF (Rosenfeld et al. 2002). In this way the concentration of the TF quickly increases and reaches a steady state after passing the repression threshold. The robustness conferred by negative feedback loops has been demonstrated in *E.coli* (Becskei and Serrano 2000). Here, transcriptional regulation of the tetracyclin repressor fused with GFP depended on lambda and two tetracycline operators bound by the tetracycline repressor. This artificial negative feedback regulation showed how this motif can stably reduce the variability and noise of transcription and the steady state levels of the fusion protein (Becskei and Serrano 2000). Furthermore, the well characterized *E. coli* chemotaxis protein network has been extensively used to characterize the robustness of networks. In this context, variation in the concentration of chemicals was shown to change the post-translational state of proteins, however the network could readapte to normal values after time (Alon et al. 1999). Interestingly, large variations in the network proteins concentration do not affect the response itself but the time in which the system could respond to the disturbance (Alon et al. 1999). In this context, negative feedback loops play a fundamental role in conferring robustness of reactions to chemotactic stimuli (Alon et al. 1999; Barkai and Leibler 1997).

A positive feedback loop occurs when a member of a network is able to positively regulate its own activity. In transcriptional networks, the kinetics of this motif is slower than in network with no feedback loops (Maeda and Sano 2006). However, despite this, if the expression levels of the TF are high, positive feedback is important for the differentiation of cells and for the maintainance of memory in an expression system (Alon 2007).

In *D. melanogaster*, the genes involved in anterior-posterior segmentation of the embryo refine and maintain their expression through a series of cross-regulatory interactions. Using a computational simulation, the affected of alterating positive autoregulatory loops for *wingless* and *engrail* was assessed. In the absence of positive feedback regulation, normal patterning could not be reproduced, while expression variation was increased. The introduction of a positive feed back motifs for these genes was sufficient to confer robustness to the network (Dassow et al. 2000). Furthermore the network, where these two genes operate, is very robust to variations in concentration and to gene duplications of elements within the network.

At a larger scale it has been observed how mir-7 in *D. melanogaster* confers robustness to the developmental program interacting with different pathways in feedback and feedforward loops consistent with the role of its target genes (Li et al.

2009). Thus, single or a combination of network motives are used to control and enhance robustness of many biological processes. This can be extended to larger and more complex networks organized in strong hubs.

These two last cases have been described as an example of distributed robustness, where different parts of the system contribute to the functionality but each of them independently has a different role (WAGNER 2005).

Rendundancy

Robustness of a system can be enhanced by the presence of multiple elements which can each perform the same function, so that the process can continue even when there is a failure of one of them: redundancy. Biologists have more than one definition for this term: (1) a system part can be removed without affecting key system's properties (Krakauer and Plotkin 2002); (2) two parts of a system perform the same or similar tasks and for this reason can be removed (WAGNER 2005). In this context, I will use strictly the second definition.

The role of redundancy has been described mainly in the case of gene deletion in different organisms when knockout strains showed only little or no clear effect at the phenotypic level. New genes arise constantly through various mechanisms, i.e. gene duplication, exon shuffling, retroposition, transposones (reviewed in Senger et al. 2004; Long et al. 2003). In most of the cases mutations can accumulate and one of the duplicated gene ends as a non-functional gene (pseudogene). In others, though, purifying selection on one or both of these genes is more relaxed so that genetic variations can accumulate by genetic drift, be selected by natural selection and fixed (Long et al. 2003). Thus, the two duplicated genes are both functional and beneficial in certain environmental conditions. Moreover early studies in mice concluded that two duplicated genes could also be functionally redundant since they might be removed without affecting the organism (Cooke et al. 1997).

Tenascin is a large glycoprotein that is part of the extracellular matrix (ECM) that is mainly expressed in the nervous system of the mouse embryo. It interacts with several proteins of the ECM: for instance it is involved in the epithelial-mesenchymal interface formation (Saga et al. 1992). Despite the apparent fundamental role that tenascin plays during mouse development, knockout mice for this gene did not show any phenotypic defects affecting the structure of the ECM, fertility or the vitality of the organism.

In *D. melanogaster*, two close homeobox transcription factors, BarH1 and BarH2, are important for eye development. They are both co-expressed in a subset of cells in the peripheral and central nervous system and yet, deletion of one of them does not cause major phenotypic effects (Higashijima et al. 1992).

Taken together these examples show that redundant functionality, brought about by gene duplication, provides different organisms with robustness against genetic perturbation.

Using genome wide approaches, many studies tried to evaluate the effect of deletion of duplicated genes to the phenotype of different organisms. In yeast, for more than a thousand duplicated genes that have been found (Gu et al. 2003), the selective deletion of one of the two copies tended to have very little or no measurable consequence on the fitness (measured by five different parameters) in comparison to the deleterious effects of the deletion of single-copy genes. This suggested a mechanism of compensation between duplicated genes. Moreover the deletion of the higher expressed duplicated gene caused a fitness reduction, indicating that there is asymmetric functional compensation (Gu et al. 2003).

Interestingly, in mouse the same experiment of knockout of duplicated gene gave a different result, finding that there is the same level of essentiality between singletons and duplicated genes (Liao and Zhang 2007). These results might indicate a different biology between mouse and yeast or a bias due to nonsystematic available data (Liao and Zhang 2007).

The retention of duplicated genes is unexpected, since if mutations in duplicated genes could accumulate without affecting the fitness of the organism this should finally lead to their degeneration. Several studies, though, indicate that new genes, the majority of which come from duplication, quickly become essential for the survival of an organism (Langkjaer et al. 2003; Chen et al. 2010; Liao and Zhang 2007). In *D. melanogaster*, 95% of young genes (<11 My) were generated by duplication and proved to be essential as singletons (Liao and Zhang 2007; Chen et al. 2010). The lethality was high at the larval and pupae stages suggesting key roles at different developmental stages (Chen et al. 2010). Nevertheless, three different hypotheses have tried to explain the reason why paralogous genes are kept: (1) the backup hypothesis, (2) the piggyback hypothesis and (3) the model of expression reduction after gene duplication (Cooke et al. 1997; Vavouri et al. 2008; Qian et al. 2010).

The backup hypothesis is the most intuitive of the three: duplicated genes are both maintained when one of them is affected by a deleterious mutation that prevents it from performing its function. Moreover, empirical data suggests that the expression similarity of paralog genes correlates with the probability that they are both essential (Kafri et al. 2005). Furthermore, one of the paralog genes is generally expressed at higher levels in normal conditions then the other one and if it is affected by deletions, a reprogramming event occurs so that the expression of the still intact duplicated gene increases to a level that is similar to the wild type condition (Kafri et al. 2005). Some authors, however, believe that this is a simplistic hypothesis that explains only a few cases (Zhang 2012) with little theoretical and empirical support.

Vavouri and coworkers (Vavouri et al. 2008) surprisingly showed how in yeast redundant duplicates are often maintained for more than a hundred years and in several cases over one billion years of evolution. This is the case of two golgi GDPase, YND1 and GDA1, that after a billion years from the duplication are still redundant. Different theoretical models have tried to explain these results. In particular, it has been proposed that the pleiotropy of genes can be the key for the preservation of redundancy throughout evolution (Nowak et al. 1997). Since the model implies that a gene performs more than one specific function, duplicated genes may be kept because of their specific non-redundant functions so that redundancy is indirectly maintained. This model is called “piggyback” model (Vavouri et al. 2008).

Comparison of RNA-Seq data in *S. cerevisiae* and *S. pombe* have suggested that after duplication, in a special type of subfuctionalization, the two daughter genes in *S. cerevisiae* decreased their expression in comparison to the ortholog single copy gene in *S. pombe*. In this way the duplicated genes are maintained for longer as both are required to maintain the optimal level of expression (Qian et al. 2010). The decreased expression is explained by the negative epistatic effect of the duplicated genes: deletion of both genes could cause a bigger effect than the deletion of only one of them (Qian et al. 2010). Under this hypothesis the two genes are not redundant in a strict sense.

In conclusion, redundancy is a property that requires two or more genes to share a similar function that confers genetic robustness to the organism. It is interesting to note that, although gene duplication is a common phenomenon, duplication of circuits or pathways is very rare.

Modularity

Modularity is a universal property of living organisms, and has been suggested to minimize the effect of perturbation within a network by shielding other optimized functional subunits from inference (Kitano 2004).

A module can be generally defined like "networks of interacting elements behaving as relatively independent units of development or function" (Schlosser and Thieffry 2000). If the modularity of networks did not exist, even small perturbations would spread throughout the entire network resulting in unexpected outcomes (Flatt 2005). Modules are wildly observed from ontogeny to transcription to pathways (Raff and Sly 2000; Hartwell et al. 1999). Each module should be independently robust using different mechanisms, i.e. network motifs and distributed robustness and redundancy. It has been observed that mutations are more likely to affect complex rather than simple organisms, and it has been proposed that reducing the complexity of independent traits that build modules would be evolutionary advantageous (Flatt 2005). Furthermore a new function can be built in a new module using already existing structured units, a phenomenon called co-option. The signal transduction pathways involved in morphological changes between species, like *hedgehog* (*hh*) in the eyespots on butterfly wings are a good example. In *D. melanogaster* it was shown that this signaling pathway is used to organize the anterior-posterior patterning in the wing disc. This function is conserved in butterflies but a new activation of the *hh* receptor, *patched* (*ptc*), is observed in the posterior part of the wing disc as well (Keys et al. 1999). This is due to an increase in the expression of *cubitus interruptus* (*ci*), a downstream component that is activated by *ptc* and is not anymore repressed by *engrail* (*en*) (Keys et al. 1999). This study indicates that during evolution the two networks have been reused and modified to allow new functions to increase the fitness and eventually be fixed by natural selection.

Regulatory regions of genes represent an immediate and clear example of modularity. In particular genes are regulated by DNA elements with very different sizes where transcription factor binding sites are clustered: *cis*- regulatory modules (CRMs). CRMs drive the expression of their target genes in precise spatio-temporal patterns, which they are sufficient to recapitulate in transgenic embryos using a heterologous reporter gene (see below).

Gene expression regulation and development

Embryonic development is a very complex process. Its ultimate outcome is the differentiation of a single totipotent cell (the fertilized egg) into many differentiated cells forming tissues and organs. All cells share an almost identical genotype that gives rise to a variety of cellular phenotypes. The stereotypic nature of development is determined by large interconnected regulatory networks, i.e. highly regulated programs of gene expression. It depends on *cis* and *trans* elements, whose integrative action results in the precise control of gene expression, both in time and space.

Transcription is fundamental to each and every cell of an organism independent of the developmental stage or adult life. It is a process that is regulated at multiple different steps to lead a DNA dependent RNA polymerase to transcribe the information contained in the DNA to RNA. It requires that the DNA upstream the transcription start site (TSS) is accessible to allow the transcriptional machinery to assemble and for activator or repressor TFs to bind to regions that are important for gene expression regulation (promoter, proximal or distant enhancers). The immature mRNA is subjected to splicing, as well as to 5' capping and 3' cleavage and polyadenylation. All these processes are strictly regulated and important to guarantee the mRNA's stability and transport from the nucleus to the cytoplasm. In light of all of these regulatory steps in gene expression, transcriptional initiation is the first and major step in the control of spatio-temporal patterns of expression in eukaryotes thanks to several elements that, together, contribute to ensure precise patterns of gene expression.

Transcription initiation control

During the initiation of transcription, the transcription machinery is recruited and assembled upstream the TSS. This region is called the core promoter and consists of many interchangeable sequence elements that dock the pre-initiation complex (PIC). Initiation of mRNA transcription depends on the assembly of a complex containing the RNA polymerase II (RNAPII) and several general transcription factors (reviewed in Loots et al. 2005; Smale and Kadonaga 2003). For example, TFIID is a protein complex consisting, among others, of the TATA binding protein (TBP) which interacts with sequence elements in the core promoter, called the TATA box (Fig. 2). Other elements in the core promoter have been identified including the initiator elements (Inr), the downstream promoter elements (DPE), both important for the recruitment of TFIID and the B recognition element (BRE) where TFIIB binds (Smale and Baltimore 1989; Smale and Kadonaga 2003) (Fig. 2).

Comprehensive studies in *D. melanogaster* have classified different classes of promoters depending on the motifs in the core promoter and the expression of the target genes. For example, the core promoter of tissue specific genes is characterized by the presence of the TATA box and the Inr element, while developmental gene promoters have an Inr alone or in combination with DPE (Ohler 2006; Engström et al. 2007)

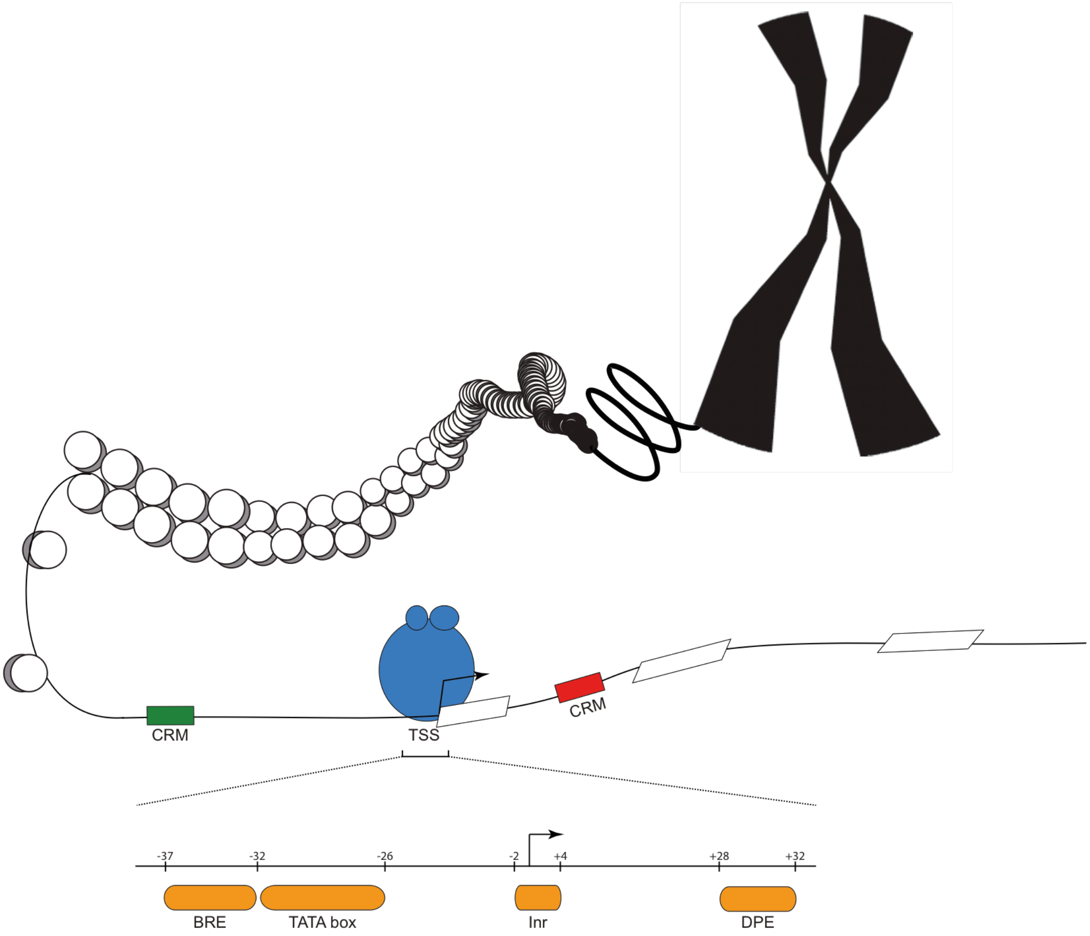


Fig. 2 Transcription initiation control. A summary of promoter elements. To recruit RNA polymerase II (RNAPII) and to activate transcription of the gene, sequence-specific regulatory proteins, TFs, bind to specific sequence patterns. The region around the TSS has several sequence patterns: the TATA box, BRE, DPE and initiator (Inr). Their location of patterns relative to the TSS is shown.

Different studies using the 5'CAGE technique in *D. melanogaster*, mouse and human have shown the existence of two different classes of promoters: (1) "broad" promoters with broad regions of distributed TSSs and (2) "peak" promoters with a single sharp TSS (Hoskins et al. 2011; Carninci et al. 2006). "Peak" promoters are enriched in TATA, Inr, and DPE elements associated with tissue specific genes, while "broad" promoters are associated with a range of less well characterized weak motifs related to housekeeping genes (Hoskins et al. 2011). "Peak" promoters are strongly and significantly associated with genes that have restricted temporal and spatial expression patterns, such as developmental genes (Hoskins et al. 2011). Furthermore, several studies showed that RNAPII accumulates immediately downstream of the TSS of housekeeping and developmental genes “broad” promoters (RNAPII pausing) (reviewed in Hong et al. 2008; Nechaev and Adelman 2011; Frankel et al. 2011).

Recent epigenetics data from different species provided an additional way to identify and classify promoters (reviewed in Ghiasvand et al. 2011; Lenhard et al. 2012; SSHASHIKANT and RUDDLE 1996).

Using chromatin profiles in a human study, the relative contribution of enhancers and promoters to the expression of different classes of genes was investigated (Ernst et al. 2011). Looking at 9 chromatin states in 9 different human cell types, a systematic comparisons of over 1000 promoters suggested that developmental genes are strongly regulated both at the promoter and the enhancer level, while tissue specific genes seemed to be predominantly dependent on enhancers since they showed less diverse promoter states (Ernst et al. 2011).

Transcription factors

A transcription factor (TF) is a protein that contains both a DNA-binding domain, which allows the protein to bind to short degenerated DNA sequences or motifs, and an activator or repressor domains, through which they can positively or negatively regulate the expression of a gene via protein-protein interactions with chromatin modifying enzymes and/or the basal transcriptional machinery. Each TF has a certain specificity for a DNA motif: some of them bind to very strict sequence motifs, others are much more flexible. For instance Reb1 plays an important role in transcriptional regulation in yeast. Chip-exo data showed that it recognizes a TTACCCG consensus sequence with high affinity and even a single-nucleotide deviation from this motif is associated with lower affinity (Rhee and Pugh 2011). Conversely Phd1, another TF involved in yeast pseudohyphal growth, binds to five recognition sites although they have distinct sequences in several positions. This does not affect the affinity of the TF, it rather reveals its flexibility in DNA recognition (Rhee and Pugh 2011).

More complex mechanisms, in addition to the simple recognition of DNA motifs are also involved in the control of enhancer occupancy and in the regulation of gene expression. Indeed, the ability of TFs to form homo-, hetero- or multimers can influence their specificity. TFs bind to small regions of the genome where TF binding sites are clustered: the importance of their combinatorial binding was shown in different studies. For instance, during *D. melanogaster* development, the striped expression of the pair-rule gene *even-skipped* (*eve)* depends on the interplay between activators and repressors (Arnosti et al. 1996b). The recruitment and combinatorial binding of broadly expressed activators, such as *bicoid* and *hunchback*, in combination with the restricted expression of repressors, like *Kruppel* and *giant*, have been proved to be important for the precise boundaries of the second stripe (Stanojević et al. 1991a).

In other cases, the expression of a gene depends on the synergistic and antagonistic interaction between TFs in different tissues. An example of synergistic interaction is represented by pMad. It is an effector of the decapentaplegic (DPP) signaling pathway that plays and important role in the induction of cardiac, visceral and somatic mesoderm differentiation after gastrulation (Frasch 1995). In the cardiac mesoderm, expression of the homeobox gene *tinman* (*tin*) depends on the TinD enhancer, where pMad binds together with the Tin protein, in a positive feedback loop (Yin et al. 1997; Xu et al. 1998). Mutation of the tin binding sites within this enhancer suggested that pMad alone is not sufficient to activate its expression (Xu et al. 1998): however cooperative binding of pMad with the Tin protein is essential to trigger full levels of mesodermal *tinman* expression (Xu et al. 1998). Furthermore combinatorial binding between Tin, pMad and other three TFs has been shown to be extensively used during the differentiation of cardioblasts (Junion et al. 2012). In other cases, TFs can either have an activator or a repressor activity depending on the other TFs with which they interact. For example *Twist* (*twi*) is a basic Helix Loop Helix (bHLH) TF important for mesoderm gastrulation and its subsequent specification in *D. melanogaster* (Leptin 1991). Twist homodimers function as an activator in the early mesoderm and in the somatic mesoderm whereas Twist works as repressor when it forms heterodimers with a ubiquitously expressed bHLH transcription factor, *Daughterless* (*Da*). *In vivo,* over-expression of a Twi-Da heterodimer caused a very strong muscle phenotype, indicating that Da-Twi heterodimers compete with Twi-Twi homodimers for occupancy to mesodermal enhancers and therefore disrupts mesoderm development (Castanon et al. 2001).

ChIP-chip and ChIP-seq data indicate that TFs bind to several thousands of locations genome-wide and that the overall binding landscape can change at different developmental stages or in different treatment conditions. In particular, looking at the binding of TFs involved in mesodermal specification and differentiation at several developmental stages of the *D. melanogaster* embryonic development, the occupancy data reflected temporal progression, cell lineage identity or activation upon specific stimulation. The same TF can bind to some enhancers continuously through multiple stages of development, while it binds to other enhancers in a transient and stage dependent manner (Sandmann et al. 2007; 2006; Jakobsen et al. 2007).

TF binding is also influenced by nucleosomes positioning at enhancers. Indeed, in both *D. melanogaster* and humans, nucleosome free regions are tightly linked to DNA accessibility in chromatin (Li et al. 2011; Degner et al. 2012). Some TFs called pioneer factors have the ability to bind to inaccessible, or nucleosome-bound, DNA and recruit chromatin remodeling factors that lead to nucleosomes repositioning and thereby local opening of the chromatin (Biddie et al. 2011).

Taken together, these data indicate that TF occupancy is controlled through the interplay of DNA binding specificity, the availability of combinatorial partners and cofactor and Nuclesome positioning.

cis Regulatory Modules

Transcription factor recognition sites are clustered in regions of the genome, called enhancer elements or *cis* regulatory modules (CRMs), which are essential for the control of gene expression. CRMs are short regulatory elements, typically a few hundreds base pairs, driving a particular aspect of a gene expression independently of their orientation relative to the TSS (Fig. 4)(Banerji et al. 1981). CRMs can be found at large distances (distal elements) from their target genes or in introns and promoters (proximal elements) (Buecker and Wysocka 2012). This role distinguishes them from basal promoter elements that, as described, recruit the transcription machinery at the TSS of a gene and determine the site of transcription initiation (Smale and Kadonaga 2003). Thus, CRMs can be considered as a platform where different pieces of information, related to the binding of TFs to specific motifs and to their function as activators and repressors, are integrated (Alatalo and Moreno 1987; Ip et al. 1991; Levine 2010; ALATALO et al. 1985; HOGSTAD 1978). This integration results helps define precise regulation of gene expression in specific tissues and temporal stages of an organism's life.

Many studies have investigated the underlying 'rules' that lay behind this integration and have provided valuable insights into how CRMs function. Elegant work in *D. melanogaster* tested the well-characterized *sparkling* (*spa*) enhancer analyzing its structure and function. The *spa* enhancer is important for the expression of the *dPax2* gene, which is sufficient to specify cone cell fate and to integrate the information from Notch and EGFR/MAPK pathways (Swanson et al. 2010). Mutagenesis analysis showed that the enhancer is densely packed with regulatory sites and shuffling the TF binding sites randomly across the enhancer sequence showed a significant switch in cell type specificity or a decreased activity of the enhancer, suggesting that the configuration of motifs plays an important role in its function (Swanson et al. 2010).

In another case, the enhancer of the virus-inducible interferon β gene requires the coordinated activation and binding of the ATF-2/c-Jun, IRF-3, IRF-7 and NFκB (i.e. p50/RelA) TFs in a nucleosome free region between -102 and -47 bp from the TSS of the gene (Agalioti et al. 2000). This 55 bp stretch of DNA was subdivided in four regulatory domains where 8 individual TF binding sites were identified (Fig. 3) (Panne et al. 2007). Single point mutations in any one of the 8 TF binding sites caused malfunction of the enhancer. Moreover, the insertion of 6 bps between TF motifs drastically reduced the activity of the enhancer, which could be reestablished when the relative TF position faced the same side of the helix (Thanos and Maniatis 1995). This suggested that the relative position of TF binding sites are important to ensure correct protein-protein interactions that facilitate cooperative TF recruitment (Thanos and Maniatis 1995; Panne et al. 2007). Moreover the composite model showed that interactions between the adjacent DNA binding domains of the 8 proteins at the enhancer created a continuous surface that underlays its cooperative occupancy (Panne et al. 2007). After the cooperative TF binding, the complex is stabilized by a subsequent recruitment of CBP/p300 that interacts with all the TFs through different protein domains (Fig. 4) (Panne et al. 2007).

Several other studies have investigated the relative position of motifs and the cooperative recruitment of TFs on enhancers. In *D. melanogaster* the innate immune response results in the activation of two classes of TFs: the Rel/NF- B family of transcription factors (i.e. *Dorsal*) and the GATA family of zinc finger transcription factors (i.e. *Serpent*). It was observed that half of the most strongly activated immunity genes respond to the synergistic binding of these factors 200- 300 bp from the TSS (Senger et al. 2004). Moreover, they often bind to sites that are only 50 bp away and are in the same relative orientation in the double helix. Mutations of this structure have significant repercussion on the activity of CRMs (Senger et al. 2004).

These studies suggested that motif positioning such as the order, the orientation and the spacing, ensures the correct position of TFs facilitating their protein-protein interactions essential for cooperative binding. This has led to the enhanceosome model of enhancers (Merika and Thanos 2001). The model features a high degree of cooperativity between enhancer-bound proteins, such that alteration in individual binding site posititioning can have drastic effects on enhancer output. “The function of the enhanceosome is thus more than the sum of individual transcription factor contributions, but emerges from a network of interactions” (Arnosti and Kulkarni 2005). This cooperative binding is associated with sharp activation of the CRM’s activity: an on/off output that may be essential for rapid response to appropriate stimuli (Fig. 4) (Kulkarni and Arnosti 2003).

Not all the enhancers have the strict organization described above. Indeed, many developmental enhancers display no or much looser architectural constraints. In particular, the regulation of developmental genes’ expression depends on CRMs that can have both additive and cooperative inputs to regulated expression in the same cell at a given time. In a synthetic study, activator and short-range repressor sites were put together with a 100 bp distance between each other. This compact element had distinct adjacent sites representing active and inactive state at the same time (Kulkarni and Arnosti 2003). In this case, co-expression of repressor and activator TFs in the same cell indicated that transcription is driven by one cluster of activators while another cluster of motifs in the same element is turned off by repressors (Kulkarni and Arnosti 2003). Furthermore, the transcriptional expression level varies with the number of activator sites given a fixed number of repressive sites suggesting an additive effect on the gene expression (Fig. 3) (Kulkarni and Arnosti 2003). All together these data suggested that the enhanceosome model does not explain the activity of this class of enhancers. It is instead described by a "billboard" model that allows a more flexible position of motifs where some TFs can still bind in a cooperative way, while other factors bind to the same enhancer in an additive manner. Moreover in this case the basal transcriptional machinery plays an active role in interpreting signals presented by the enhancers, analogous to information display.

Other more recent evidences suggest that a third model might explain other classes of enhancers. As previously mentioned, when reshuffling the same TF binding sites of a regulatory region, different patterns of expression in comparison to the original sequence could be observed (Swanson et al. 2010). Besides, different motifs at the enhancer level could result in a very similar expression activity (Zinzen et al. 2009). In addition in *D. melanogaster* the differentiation of cardioblasts depends on 5 transcription factors: *tinman* (*tin*), *dorsocross* (*doc*), *pannier* (*pnr*), pMAD (effector of the Wingless pathway) and dTCF (effector of the Hedgehog pathway) (Reim and Frasch 2010). A recent study showed that those 5 TFs bind together to a large set of enhancers with a great diversity in terms of their TF motifs (Fig. 3), with some enhancer not even having the presence of motifs for all five factors. Despite this, *in vitro* experiments showed that when one of the TFs is removed the activity of the CRM is drastically compromised. This model suggests once there is motifs present for at least three TFs, it creates enough of a surface interface to allow all five TFs to bind via protein-protein interaction, called a "TF collective" (Junion et al. 2012).

As already discussed, CRMs drive some aspect of the complex expression of its target gene. The spatio-temporal expression of a gene is explained by the combination of all the CRMs acting on it, reflecting the transient presence of particular TFs. For example, the regulation of the expression of the *tin* gene depends on three different CRMs: tinA-B, tinD and tinC enhancers that in a transgenic embryos each regulate part of the target gene's expression; in the the early mesoderm (embryonic stage 8), in the dorsal mesoderm (embryonic stage 10) and in cardioblast (embryonic stage 16), respectively (Zaffran et al. 2006). Thus they regulate *tin* expression at different developmental stages and in different tissues as result of their different TFs occupancy profiles (Zaffran et al. 2006; Reim et al. 2005; Reim and Frasch 2010; Yin et al. 1997; Junion et al. 2012).

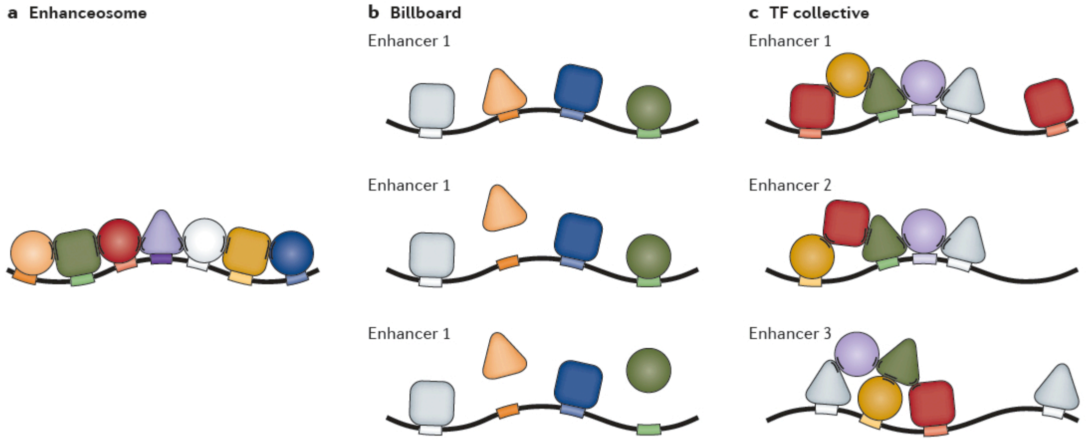


Fig. 3 Model of enhancer activity. A. The enhanceosome model: the cooperative binding of all transcription factors (TFs) to an enhancer is essential for activation of the enhancer. **B.** The billboard model: the positioning of TF binding sites is flexible and subject to loose organizational constraints. **C.** The TF collective: a same set of TFs bind to many enhancers occupying each one of these enhancers in a different manner. Image from (Spitz and Furlong 2012).

Another example is represented by *eve*. Five separate CRMs located upstream and downstream of the transcription unit regulates the expression of *eve* in different stripes and in different tissues later in the development. Each of them follows the same basic logic previously described for the stripe 2 enhancer (Jaynes and Fujioka 2004).

Chromatin loops are important for the general structure of chromatin and also for enhancer-promoter interactions that regulate the expression of a single gene (Dekker et al. 2013; Dixon et al. 2012; Sexton et al. 2012; Noordermeer et al. 2011). Once TFs bind to an enhancer, chromatin loops are thought to bring the active enhancer in close proximity to the promoter region of the target gene. One of the best studied examples is represented by the β-globin locus where the locus control region interacts with the β-globin gene located 40-80 kb away, through a specific a transcription factor GATA1 (Deng et al. 2012). *sonic hedgehog* (*shh*), a gene important for the patterning of the neural tube and limbs (Vokes et al. 2008), is another example of long-range enhancer-promoter interaction in mouse. Its expression in the neural tube depends on two regulatory regions located relatively close to the TSS of the gene while the expression of *shh* in the limbs is regulated by an enhancer located almost one megabase away (Amano et al. 2009).

All together these examples indicate that, especially during development, the activity of a single enhancer requires the combinatorial information of TFs expressed or activated at different times in different cells of different tissues which are able to bind to motifs of these regulatory regions. Moreover, they also demonstrate that the overall spatio-temporal expression of a developmental gene depends on the interplay of different CRMs.

Redundancy and transcriptional regulation

In 1942 Waddington proposed that developmental reactions are canalized against environmental perturbation (Alatalo and Moreno 1987; Waddington 1942). In order to be robust, an organism has to develop several strategies, the molecular components of which were subsequently shown to include modularity, network motifs and redundancy (see ‘Mechanisms of robustness’).

Redundancy can be mainly explained by duplicated genes. However, redundant interactions in a developmental gene regulatory network were observed as well. The specification of intestinal cells in *C. elegans* relies on a transcriptional cascade that involves the activation of the *ent-2* gene by two transcription factors: *end-1* and *end-3* (Chapman et al. 2000; Raj et al. 2010). The deletion of one of these TFs caused a lost of intestinal cells in 5% of mutants explained by an increase of the noise in the system (FEINSINGER and SWARM 1982; Raj et al. 2010). Such results suggested that the expression of these two TFs help to buffer the system from stochastic variability.

In *D. melanogaster*, precise developmental gene expression can arise from stochastic transcriptional activity (FEINSINGER and SWARM 1982; Little et al. 2013) since transcribed loci have 45% intrinsic noise that is independent of any specific promoter-enhancer architecture. The precision is achieved for example through spatio-temporal integration derived from the accumulation and diffusion of mRNA (FEINSINGER and SWARM 1982; Little et al. 2013).

The complex spatio-temporal expression of a gene is the result of the combined activity of individual enhancers, whose deletion can lead to severe phenotypes in many organisms. In humans a number of pathologies are associated with deletion or mutation in distal regulatory region (Schluter and McPhail 1992; Kleinjan and van Heyningen 2005). For example, the Van Buchem disease is a homozygous recessive disorder characterized by an accumulation of bone mass that gives rise to facial distortions, enlargement of the mandible and the head and entrapment of the cranial nerves (McPhail 1992; Loots et al. 2005). This pathology is associated with a structural variation downstream the SOST gene that affects one of its enhancers driving the expression of the gene in bones (Day et al. 1994; Loots et al. 2005). In other cases, though, deletion of enhancers did not cause a clearly recognizable phenotype implying the presence of redundant mechanisms, i.e. similar modules that can replace each other when one fails. For instance, *Prx1* is a gene important to promote limb skeletal elongation and *Prx1* null mutants die at birth with significantly shortened limbs and craniofacial defects (Day et al. 1994; Cretekos et al. 2008). Surprisingly, homozygous deletion of a conserved enhancer did not cause any detectable phenotypic effects (Day et al. 1994; Cretekos et al. 2008). These results suggest that in the *Prx1* locus, one or more regulatory elements compensate for the deletion of this enhancer, at least in lab conditions.

Redundancy can also occur within an enhancer, where multiple redundant TF binding sites contribute to the stability of the expression of the target gene under different environmental conditions. This is the case of the *eve* stripe 2 element (S2E). The S2E contains 12 transcription factor-binding sites, including activator sites (for Bicoid and Hunchback) and six repressor sites (for Giant and Kruppel) that determine *eve* expression in the anterior half of the embryo, between the Giant and Kruppel domains (Pigliucci et al. 2006; Stanojević et al. 1991b; Arnosti et al. 1996a). Removal of 200 bp of the S2E did not affect the expression of the second stripe under laboratory conditions concluding that this part of the enhancer and the two Kruppel binding sites are not functional (Small et al. 1992; De Jong 1995; Via et al. 1995; Pigliucci et al. 2006). However, under stressful conditions, the minimal *eve* stripe 2 element was not able to rescue the *eve* null mutant, proving the presence of redundant binding sites that confer robustness to the expression of the target gene (Ludwig et al. 2011; de Jong 2005).

In addition to intra-enhancer redundancy, there is extensive literature in vertebrates and invertebrates of what suggests inter-enhancer redundancy; two or more enhancers driving the expression of the same target gene in a similar spatio-temporal way, regulate the transcription of developmental genes (Fig. 4) (Hong et al. 2008; de Jong 2005; Frankel et al. 2011). In some cases the redundancy was evaluated in a reporter assay that suggested the potential redundancy among regulatory elements (SSHASHIKANT and RUDDLE 1996; Pigliucci et al. 2006). For example in *D. melanogaster* Sgs-4 is one of the glue genes that codes for a glycoprotein that is secreted in the lumen of the salivary gland and expelled during the end of the third instar for puparial adhesion (Korge 1975; Waddington 1953). Transcriptional regulation of this gene requires three different elements that in a transient transformation assay showed overlapping expression (Jongens et al. 1988; Pigliucci et al. 2006; Waddington 1953). Here redundancy in the strict sense of its definition was not proven, since experiments that assessed the effect of a deletion were not performed. In other cases, clear redundancy or partial redundancy was more accurately investigated. One example is *dac*, the most downstream member of the *D. melanogaster* retinal determination network, and therefore essential for the eye development. Indeed, null mutants show defects in leg development and a complete absence of eyes (Pappu et al. 2005; Waddington 1953). Dac expression depends on two enhancers, one at the 3’ and one located in the eighth intron of the gene. However the deletion of one of these causes only moderate phenotypic defect suggesting that the two are partially redundant (Pappu et al. 2005; Waddington 1953). Another example is the *Hoxd* gene, which is responsible for the development of both the proximal and distal limb segments in mouse. The expression of this gene depends on the activity of several regulatory regions separated by large distances (Montavon et al. 2011; Waddington 1953). Chromatin conformation captured experiments showed that those regions come together to regulate the same gene in the same cells and deletion experiment proved that they act in a partial redundant way (Montavon et al. 2011; Waddington 1953).

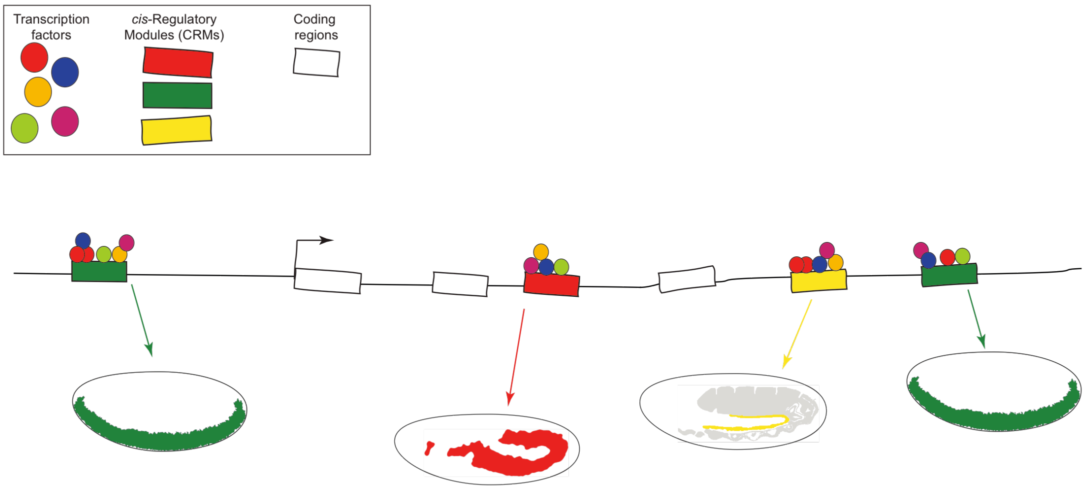


Fig. 4 Tissue-specific gene expression is driven by the activity of *cis*-regulatory modules. A locus of a hypothetical gene whose expression is driven by four *cis*-regulatory modules with defined spatio-temporal activity in the *Drosophila* embryo. Each CRM consists of clusters of binding sites recognized by distinct transcription factors, TFs. Note that two CRMs are active in a similar way early in development in the mesoderm regions.

In *D. melonogaster* many redundant enhancers have been described to regulate the expression of several genes at different stages of development (Perry et al. 2010; Williams 1966; Perry et al. 2012; Hong et al. 2008; Frankel et al. 2010). These multiple enhancers were identified because of their similar TF occupancy profile. In particular, primary and secondary enhancers were distinguished and, in a reporter assay, they drove the expression of the reporter gene at the same or similar way. In general, the primary enhancer localizes closer to the TSS of the target gene than the secondary or shadow enhancer that is located several kb away, in remote intergenic regions or in introns of other genes (Hong et al. 2008; Williams 1966; Eshel and Matessi 1998). The evolutionary pressure on those enhancers could be different: the primary enhancers could be more constrained than the shadow ones which, in turn, could accumulate more mutations and evolve differently (Hong et al. 2008; West-Eberhard 2003).

One of the first examples of redundant enhancers with these properties was described in *D. melanogaster*. *brinker* (*brk*) is a transcriptional repressor acting downstream the Dpp pathway (Jaźwińska et al. 1999; Eshel and Matessi 1998). In early embryonic stages, its expression depends on two different CRMs located at the 5' (primary enhancer) and 3' (shadow enhancer) of the gene in the intronic region of the neighbor gene *Atg5* (Hong et al. 2008; Losos et al. 2000). In a reporter assay, they share similar activity at stage 5 of *D. melanogaster* embryonic development and for this reason they were defined redundant. A later paper confirmed the overlapping expression of the two CRMs at early stage but excluded the redundant function (Dunipace et al. 2013; Losos et al. 2000). Using a BAC strategy, they hypothesized sequential recruitment at the promoter and, as consequence, sequential regulation mediated by *brk* itself (Dunipace et al. 2013; Losos et al. 2000).

Another study showed that the regulation of the *snail* gene depends on redundant enhancers. *snail* (*sna*) is a TF that together with *twist* (*twi*) is responsible for the gastrulation of the mesoderm. Its expression is regulated by a proximal enhancer located upstream of the transcription starting site (TSS) and by a recently identified enhancer, defined redundant, located in the first intron of a neighboring gene (Perry et al. 2010; Gurevitch 1992). Quantitative confocal imaging assay, BAC transgenesis and stress-induced experiments suggested that the two enhancers ensure reliable activation of the snail expression (Perry et al. 2010; Gurevitch 1992). The deletion of the primary or of the shadow enhancer caused increased failure rate of the gene transcription in the entire domain of expression. A subsequent study, instead, claimed that the two enhancers were not redundant since they did not share the same properties in endogenous conditions (Dunipace et al. 2011; Gurevitch 1992). In particular, they showed that a deletion of the distal enhancer was not able to rescue the phenotype of *sna* mutants while a deletion of the proximal enhancer did not increase the viability of mutant embryos except in extreme environmental condition. Moreover, those two enhancers have slightly different *cis*-regulatory logic, which support distinct expression but together ensure the expression pattern precision of the gene.

In humans, the deletion of a regulatory region 20 kb upstream the ATOH7 gene is responsible for the nonsyndromic congenital retinal nonattachment caused by defects in the retinal ganglion cell (RGC) and optic nerve development (Ghiasvand et al. 2011; Day et al. 1994). A reporter assay showed that this enhancer is partially redundant with a proximal enhancer, closer to the gene. It was speculated that the presence of redundant enhancers is important for the expression level of the protein: it could boost the expression of the gene up to a certain threshold that is required for proper eye development. The deletion of one of the redundant enhancers could therefore lead to the pathology due to a failure to reach the appropriate mRNA levels and protein concentration (Ghiasvand et al. 2011; Waddington 1942).

An elegant study investigated the role of redundancy in *D. melanogaster* in the regulation of *shavenbaby* (*svb*) expression (Frankel et al. 2010; Waddington 1942). *svb* is a gene encoding a transcription factor that directs development of the *Drosophila* larval trichomes, whose transcriptional regulation depends on 5 different enhancers at the 5' region of the gene (Frankel et al. 2010; RENDEL and SHELDON 1960). The enhancers drive expression in extensive overlapping regions, indicating that redundant enhancers partially share their territory of expression at the same developmental stage (Frankel et al. 2010; RENDEL and SHELDON 1960). In ideal environmental conditions, deletion of one of these enhancers did not cause phenotypic changes in the bristle number, however bristle number varied when embryos were grew at extreme temperature. Thus, redundant enhancers are important to ensure the expression robustness and to guarantee the patterning precision of developmental genes.

Aim of the project

So far, all the studies aiming to investigate the role of redundant enhancers in transcriptional regulation have focused on one single locus of a genome and have tried to generalize their findings to all possible cases.

The aim of the project is to globally identify putative redundant enhancers and observe how general is this phenomenon in a developing organism, investigating directly the intrinsic properties and evolutionary forces that characterize and maintain PREEs.

We used two complementary approaches to obtain a global map of putative redundant enhancers: those with very similar TF occupancy and elements with similar temporal and spatial activity. Here, our definition of redundancy is enhancers that have redundant activity in some or all cells where they are active. We have focused on a set of ChIP defined *cis*-regulatory modules (ChIP-CRMs) representing the occupancy of five TFs across multiple consecutive developmental stages (Zinzen et al. 2009; RENDEL 1959). As these TFs have overlapping expression and form part of an interconnected network regulating mesoderm specification, Zinzen et al (Zinzen et al. 2009; Polaczyk et al. 1998) grouped single ChIP-peaks in close proximity to define 8008 ChIP-CRMs that are bound by one or factor. Importantly, 97% of these ChIP-CRMs function as developmental enhancers when tested in vivo, using transgenic reporter assays (Zinzen et al. 2009; Polaczyk et al. 1998). We have used transcription factor (TF) occupancy data of five key mesodermal transcription factors and the predicted activity of 8008 mesodermal enhancers to identify across the genome potential partially redundant enhancers (PREEs) that have a similar TF binding profile or share similar predicted activity.

Moreover we have generated a list of structural variations from 162 wild type inbred and fully sequenced D. melanogaster lines that are part of the Drosophila Genetic Reference Panel (Mackay et al. 2012; Polaczyk et al. 1998). Having a list of PREEs and genomic deletions of a wild type population, we could observe the direct effect of deleted PPREs on the expression of their target genes and evaluate the general role of redundant enhancers in conferring robustness to the transcriptional regulation of a developing organism.

Thus, we defined as PREEs those regulatory regions that control the expression of target gene in a similar spatio-temporal way and, if individually deleted under standard laboratory conditions do not impair the development of the organism, in agreement with the definition of redundancy (WAGNER 2005; Wagner and Altenberg 1996). PREEs, located at any distance from the TSS of the target gene, do not have the same identical properties but they share large or small overlapping regions of activity for some time during the development. Furthermore we do not distinguish between primary and secondary enhancers, since we want to avoid giving them implicit and improper degrees of importance.

Results

Identification of partially redundant enhancers

To generate a genome-wide map of putative redundant enhancers, we used two criteria for their definition: (1) the similar activity and (2) similar TF binding profile.

(1) In previous approaches that attempted to define redundant enhancers, the similarity of the transcription factor occupancy profile was used as criterion (Hong et al. 2008; Wagner et al. 1997). Thus, we investigated the binding of five transcription factors fundamental for the specification and differentiation of the mesoderm in *D. melanogaster* at different stages of development (Sandmann et al. 2006; Eshel and Matessi 1998; Sandmann et al. 2007; Jakobsen et al. 2007; Liu et al. 2009; Zinzen et al. 2009) (Fig. 5). We took advantage of a complex matrix of 15 different entries represented by ChIP-chip data on five transcription factors (twi, tin, Mef2, bin, bap) at several developmental stages to investigate similarity in the TF binding profile. Here similarity is defined by the rank based spearman correlation on the peak intensity for each TF across 8008 mesodermal ChIP-CRMs within a 50kb distance of each other (Zinzen et al. 2009; Clarke and McKenzie 1987). Fig. 5 shows a real case where the TF profile of the two enhancers follows the same binding trend across the development. We identified 609 unique pairs of putative redundant enhancers with distance between 0.2 and 50 kb (Suppl. Table 3).

(2) Although enhancers bound by the same combination of TFs often give rise to similar patterns of expression, a number of studies have shown that enhancers with diverse patterns of TF occupancy can also give rise to highly similar spatio-temporal activity, due to the result of transcription factor combinatorial binding (Zinzen et al. 2009; Clarke and McKenzie 1987; Brown et al. 2007). Thus, we looked for putative redundant enhancers in view of their mesodermal predicted activity. The activity of 8008 CRMs was predicted in previous studies (Zinzen et al. 2009; Wagner et al. 1997) with a support vector machine approach and with ChIP-chip data for the same 5 mesodermal transcription factors (Fig. 6). They described 5 different exclusive classes depending on the expression domain (Fig. 6). We defined putative redundant enhancers those regulatory regions with a similar predicted activity that is to say a SVM specificity score > 95% for the same class of predicted activity. Using this approach we identified 1708 unique pairs of PREEs (Fig. 6).

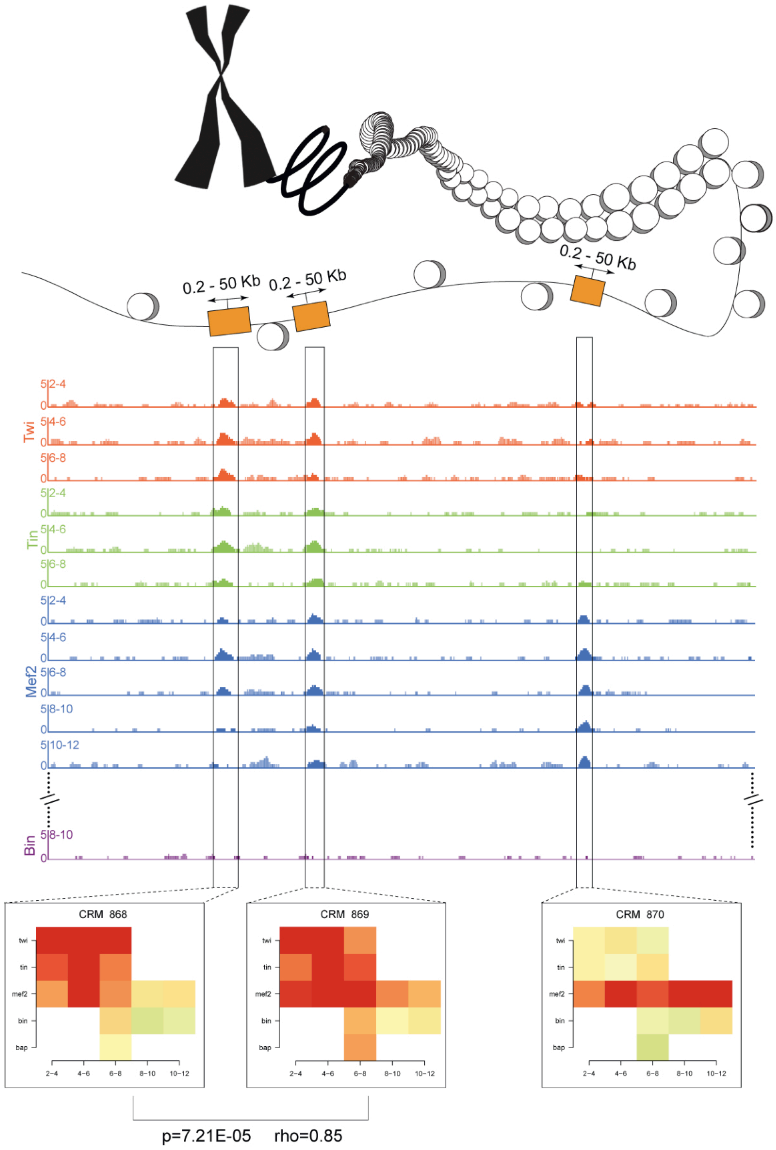


Fig. 5 Transcription factor binding profile similarity for PREEs' prediction.

We investigated the TF binding profile of five transcription factors important for the mesodermal development at different time points, data that have been used to define mesodermal CRMs (orange)(Zinzen et al. 2009; Clarke and McKenzie 1987). ChIP-chip data for Twist (red), Tin (green), Mef2 (blue), Bin (purple) are shown. Here is a real case in which two CRMs (CRM868 and CRM869) in a 50Kb window are predicted to be PREEs by the spearman correlation on ChIP signals (rho≥0.8). Visualizations of this similarity are the heatmaps (bottom) of all mesodermal TF binding signals at different time points during embryonic development for each CRM (i.e. compare CRM868 vs CRM869 or CRM868 vs CRM870).

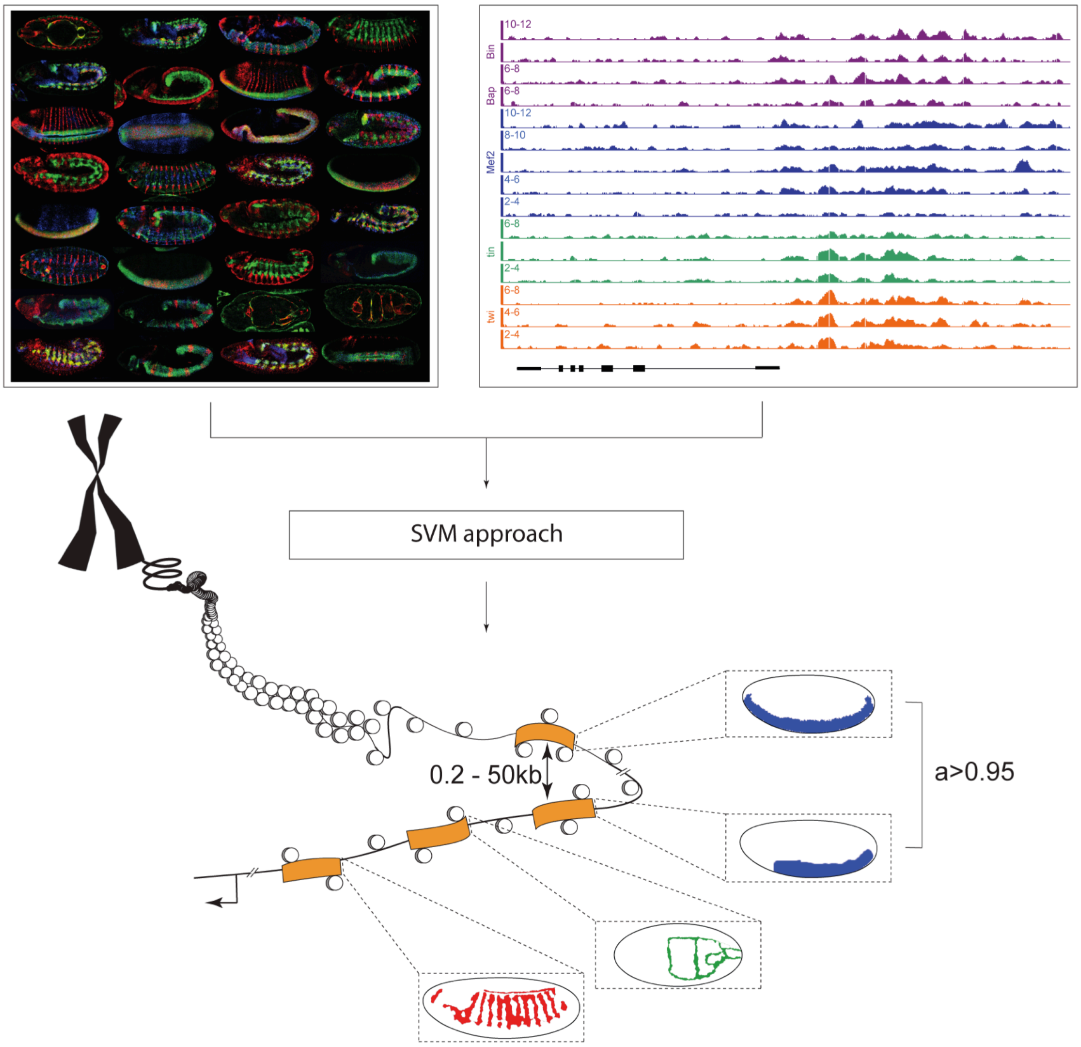


Fig.6 Similar spatio-temporal enhancer activity in the prediction of PREEs.

The activity of 8008 mesodermal enhancers was predicted using a support vector machine (SVM) approach considering mesodermal TF occupancy data (Zinzen et al. 2009; Eshel and Matessi 1998). To each CRM, a probability score of activity in 5 classes is assigned: early unspecified mesoderm (blue), visceral mesoderm (green), somatic mesoderm (red) and two more complicated classes: early mesoderm and somatic muscle or visceral muscle and somatic muscle. In this analysis we defined PREEs, CRMs in a 50 kb distance with a SVM specificity score (s) higher than 0.95 in the same class of expression.

The combination of these two approaches identified 2246 high confident pairs of putative redundant enhancers in 50kb windows. Interestingly, we observed that 71% of loci have minimal pairs of putative redundant enhancers while the other 29% of gene have more complex transcriptional regulation that depends on either three, four or in few cases even eight PREEs (Fig. 7). All together these data indicate that our current view of potential redundant enhancer is too simplistic.

Properties of putative redundant enhancers

The two approaches that we used for the detection of PREEs allowed us to identify 2246 pairs of putative redundant enhancers in 50 kb distance. We next investigated the specific properties of PREEs that distinguish them from non-redundant CRMs. Predicted redundant enhancers tend to be closer than expected with more then 50% of the pairs localizing within 25 kb (Fig. 7A). Due to the nature of redundancy, one of the two elements with similar function can be easily lost since it does not cause phenotypic effect or quickly diverge due to different selective force on them. In a previous work, the *shadow enhancer* was proposed to be less conserved than the *primary enhancer* suggesting different evolutionary pressure between pairs (Hong et al. 2008; Wagner et al. 1997). Since we identified a large number of PREEs we could test directly this hypothesis. We used for that the PhastCons (Siepel 2005; Wagner et al. 1997) 15-way Genome Conservation Scores and classified a CRM as conserved if it has a PhastCons score above 0.9 for over 70% of its size. Indeed, we observed that putative redundant CRM pairs contain more instances where both CRMs are conserved compared to random sets (Fig. 7B). These correlations suggest that both enhancers in a pair are under similar evolutionary constraint and argue against the hypothesis that these putative redundant enhancers act as a source of evolutionary novelty (Hong et al. 2008; DUN and FRASER 1958).

It rather suggests that although enhancers may act redundantly in one condition, they are not redundant in another context: for example in a different developmental stage, tissue or environmental condition, as recently shown for the svb enhancers (Frankel et al. 2010; RENDEL and SHELDON 1960).

We then examined if the individual motifs for each TF are under selection within each set of CRMs using Tajima’s D test to characterize the frequency of SNP alleles segregating within the 162 isogenic D. melanogaster lines. Comparing putative redundant enhancers vs non-redundant enhancers we found that redundant CRMs, taken as whole regions, have slightly more negative Tajima's D values than non-redundant CRMs. This indicates that putative redundant enhancers evolve under stronger negative selection (Tajima's D= Tajima’s D = -0.81 vs. -0.74, p = 0.016) (Fig. 7D).

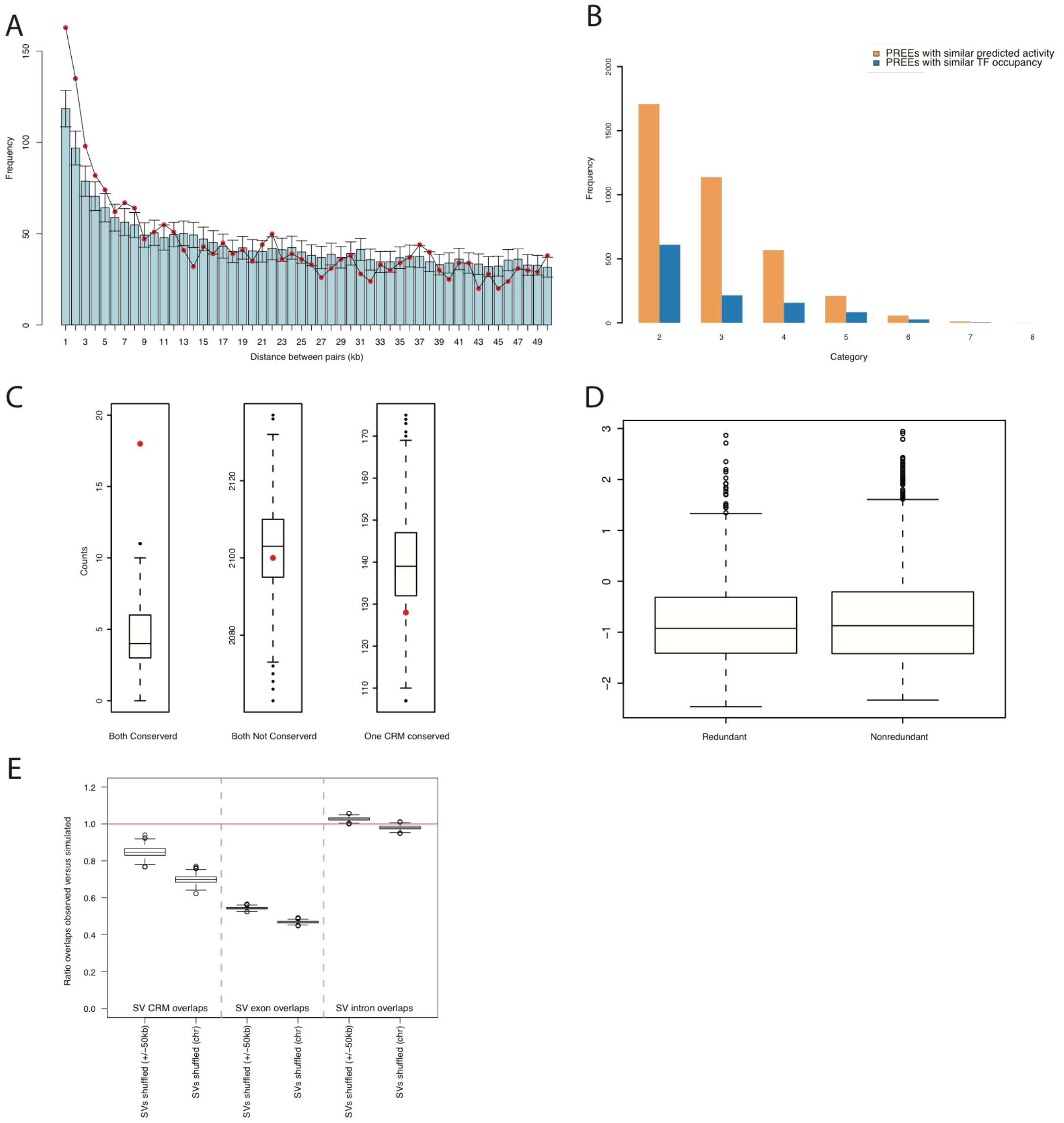


Fig. 7 PREEs properties

A. 50 kb window distance distribution of PREE pairs (red dots) vs 1000 random set of enhancers (light blue) with standard deviation. **B** Number of PREEs in 50kb window identified using similar predicted activity (orange) or similar TF occupancy profile (blue). We observe not only pairs, but set of PREEs (in some cases also eight). **C** Conservation of PREEs across Drosophilidae. We observe (red dots) significant conservation of both PREEs in a pair in comparison to random set of paired enhancers (Materials and Methods). We do not observed significant difference in those cases where one PREE of the pair in not conserved or when both are not conserved. **D** The box plot shows that the conservation of both partially redundant enhancers is confirmed by the Tajima D test. We observed a significant negative selection on paired enhancers that have been predicted to be redundant (Tajima's D = -0.81 vs. -0.74, p = 0.016). **E** Box plots showing the significant depletion of structural variations in exonic and enhancers regions in comparison to introns when compared to observed and simulated overlaps.

Moreover, we performed a Gene Ontology (GO) analysis on all the pairs of putative redundant enhancers to determine their importance in the regulation of specific classes of genes. We found a significant 40% of pairs assigned to the common target gene enriched for terms related to development, transcriptional regulation, locomotion and response to stimuli (p-value <0.05) (Suppl. table 1).

Together these data suggested that pairs of putative redundant enhancers are both selectively maintained against the theory that the single elements of the pairs are under different selective pressure. Furthermore it suggest that they could have slight different PREEs roles in the spatial and temporal gene expression and important to ensure the robust expression of genes fundamental for the development of an organism perhaps in a buffering against mutations, as opposed to having a role in evolutionary diversity.

Natural structural variations affect PREEs

Redundant, or partially redundant, enhancers can compensate for mutations that render one of the enhancers dysfunctional, as shown in the svb (Frankel et al. 2010; RENDEL and SHELDON 1960) or dac (Montavon et al. 2011; RENDEL et al. 1965) loci in D. melanogaster or the Hoxd loci in mouse (Pappu et al. 2005; RENDEL et al. 1965). If the PREEs act as true redundant enhancers, the developmental reactions should not be affected by genetic perturbation on one of them. Thus we used natural sequence variation within a wild population of *Drosophila melanogaster* to determine if enhancers within a predicted redundant pair are affected by structural variation (SV). As it often difficult to predict the effect of an individual SNP on TF occupancy (Maurano et al. 2012; Scharloo 1991; Reddy et al. 2012), we focused here on deletions (structural variations (SVs)) greater than 50bp, located in the center of the enhancer and affecting at least 25% of its size. For this we took advantage of a set of 205 fully sequenced inbred lines from the Drosophila Genetic Reference Panel (DGRP) (Mackay et al. 2012; VISSER et al. 2003).

To facilitate this study we first extended our previous SV analysis from 40 lines (Zichner et al. 2012; Eyre-Walker and Keightley 2007) to 205 lines with a few modifications. The variant discovery was performed as follows: We inferred deletions in all 205 lines using four different computational tools Pindel (Ye et al. 2009; Krakauer and Plotkin 2002), DELLY (Rausch et al. 2012; Wagner et al. 1997), Genome STRiP (Handsaker et al. 2011; Rutherford and Lindquist 1998), and CNVnator (Abyzov et al. 2011; Rutherford and Lindquist 1998). We then integrated the results by merging the individual variant predictions for all samples and the four methods generating a single variant list (Materials and Methods). To minimize false positive predictions and to increase accuracy, we used only those deletions that were between 50bp and 25kb in size and that were predicted at nucleotide resolution for further analyses. Furthermore, we removed all variants that overlapped annotated repeats by more than 90%. Based on whole-genome tiling array data that was available for six of the lines (Zichner et al. 2012; Masel and Siegal 2009), we estimated the false discovery rate of the final list to be at about 15% (Materials and Methods).

Based on these structural variants, we first examined how often SVs affect different functional parts of the genome, such as exons, introns and enhancers (Figure 7E). To assess the significant of these results we performed simulations: we randomly moved all SVs 1000 times by up to 50kb up- or downstream and reassessed the overlap with the functional elements for each iteration (Figure 7E). Overall, exons are strongly depleted in deletions when comparing the overlap in the number of observed and simulated events. Enhancers show the same trend (although weaker) emphasizing their importance in the genome. For introns no clear difference between observed and simulated overlaps was observed.

We next examined the overlap of the SVs with identified 160 cases (11%) where one of the putative redundant enhancers is deleted (72 from the PREEs based on similarity in TF occupancy, 77 based on similarity in their activity, and 11 common). The flies harboring these SV mutations are alive and viable, at least under laboratory conditions, so even when one of these developmental enhancers is deleted embryogenesis proceeds largely normally, providing evidence that the function of this enhancer must be compensated for by the second 'shadow' enhancer. An essential prerequisite of this assumption, is that both enhancers drive spatio-temporal gene expression in at least partially overlapping domains. To directly test this, we examined the activity of 7 pairs of putatively redundant enhancers in detail. We chose three loci with redundant enhancers predicted based on similarity in TF occupancy, three loci predicted by similarity in spatio-temporal activity and one common locus predicted by both approaches. We categorized these loci as (1) simply regions with only one pair of predicted redundant enhancers and (2) more complex regions, where more than a simple pair of redundant enhancers was predicted.

Assessing overlapping activity of PREEs

We validated the predicted deletions by PCR in individual isogenic lines (Suppl. Fig. 1). We then tested the activity of the putative redundant enhancers and most of the enhancers in the gene loci in a transgenic reporter assay. Enhancers were cloned upstream the reporter gene LacZ and the vector was integrated in the same location in the *D. melanogaster* genome to avoid potential differences due to the integration site. We assessed the spatio-temporal activity of CRMs during the embryogenesis by double fluorescent *in situ* hybridization for LacZ and several mesoderm markers used as mesoderm tissue reference for the different developmental stages. For two out of the seven tested loci, *atx2* and *ptc*, one of the predicted partially redundant enhancers did not have any activity in the reporter assay during the embryogenesis.

In all the remaining examined loci, we observed overlapping activity more than two enhancers. For example, in the case of *rolling pebbles* and *CG42788*, we found three predicted redundant enhancers having similar spatio-temporal activity. The *rolling* *pebbles* (*rols*) gene codes for an essential protein that is part of the protein complex established at cell contact sites between precursor cells and fusion-competent myoblasts (Rau et al. 2001; Masel and Trotter 2010). *rols* is first detected in the mesoderm at the extended germ band stage in progenitor/founder cells (stage 11). The number of *rols* expressing cells increases during germ band retraction (stage 12) in somatic and visceral mesoderm. After stage 12, the expression is restricted in a subset of somatic muscles (Rau et al. 2001; Schuster et al. 1994). *rols*-mutant embryos are characterized by many unfused myoblasts and they develop until hatching (Rau et al. 2001; Eyre-Walker and Keightley 2007) proving that *rols* is an important gene for their survival.

We observed that the regulation of this gene involves the presence of three partially redundant enhancers (Fig. 8) and that one of them is deleted in 11 out of the 162 isogenic lines. Looking at the activity of the three PREEs, we could see that all of them are active in overlapping regions at stage 11 and 12 (Fig. 8A and Suppl. Fig. 3) despite a clear difference in TF occupancy (data not shown). As predicted, they are active in somatic and visceral muscles, recapitulating as well the expression of the target gene.

In the *CG42788* locus (Fig. 8B and Suppl. Fig. 4), we observed that two of the three PREEs with similar TF binding profile, drive the expression of the reporter in the visceral muscle at the embryonic stage 15-17.

The *ade5* locus shows more complexity. Here, we identified four putative redundant enhancers with predicted activity in the somatic and visceral mesoderm and *ade5* as predicted target gene: it is involved in the *de novo* purine synthesis and mutations of the gene reduce the viability of the organism being lethal if homozygous (O'Donnell et al. 2000; WAGNER 2012). We found a deleted enhancer, CRM7490, in one viable and fertile DGRP line that has activity in the somatic and visceral mesoderm from stage 11 to stage 14. Other three PREEs (CRM7483, CRM7487/8 and CRM7489) have overlapping spatio-temporal activity with the deleted enhancer: they drive *lacZ* expression either only in the visceral mesoderm or in both the somatic and visceral mesoderm, at some or all developmental stages (Fig. 9 and Suppl. Fig. 5).

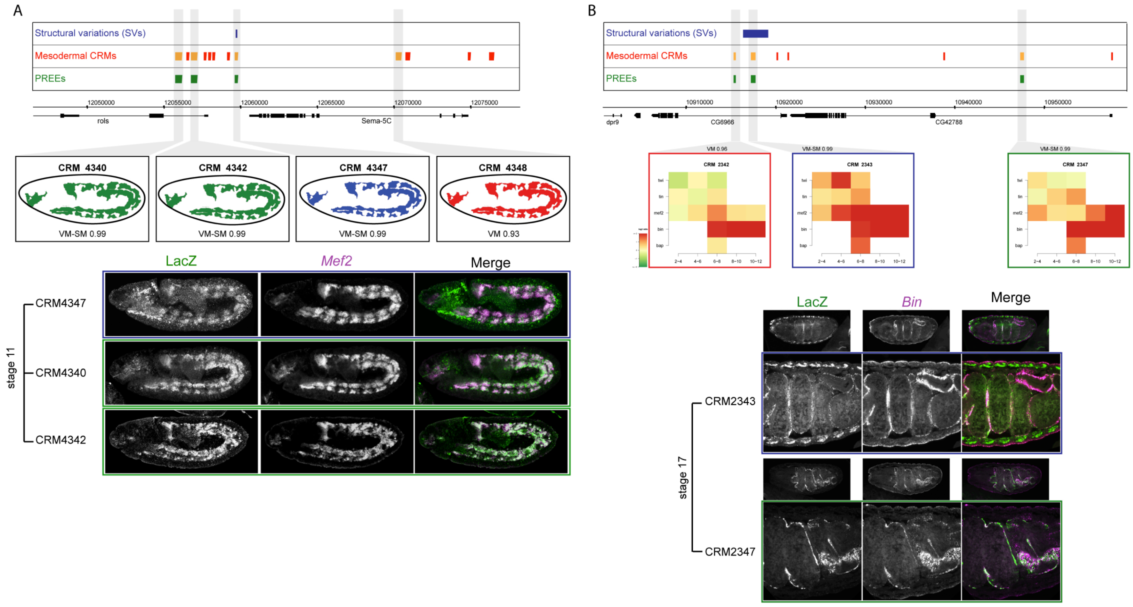


Fig. 8 Spatio-temporal activity overlap of enhancers in the rols and CG42788 loci.

A. rols locus. 3 PREEs were identified in the locus with similar predicted activity and one of them (CRM4347) affected by structural variations in several DGRP lines. SVM specificity scores and predicted expressions of the CRMs tested in a reporter assay are shown. CRM4348 (red) was included in the reporter assay since the SVM specificity score is slightly below the threshold (>0.95). However we did not observe overlapping activity with the deleted enhancer. Double fluorescent ISHs (bottom) of the lacZ reporter (green) driven by PREEs (CRM4347, 4340, 4342) and a general mesodermal marker, Mef2 (magenta), show a similar mesodermal activity of the PREEs at embryonic stage 11 (compare LacZ expression). **B**. CG42788 locus. 3 PREEs were predicted in the locus with similar TF occupancy profile and similar predicted activity. CRM2343 is completely deleted in some lines from the DGRP panel. Heatmap represent the binding of 5 transcription factors (Twi, Tin, Mef2, Bin, Bap) from stage 5 to stage 15 in 2h developmental windows (top). Despite the CRM2342 was predicted to be a PPRE, it did not share regions of overlap with the other PREEs. Double fluorescent ISHs (bottom) of the lacZ reporter (green) driven by PREEs (CRM2343, CRM2347) and a general mesodermal marker, Mef2 (magenta), show a similar activity of the PREEs in visceral mesoderm (compare LacZ expression). Despite their similarity CRM2343 shows non-overlapping activity in the somatic mesoderm. Embryo at stage 17.

All embryos are oriented anterior - left, dorsal – up. Shown Mesodermal CRMs (red), structural variation (blue), CRMs tested in reporter assay (orange) and PREEs (green).

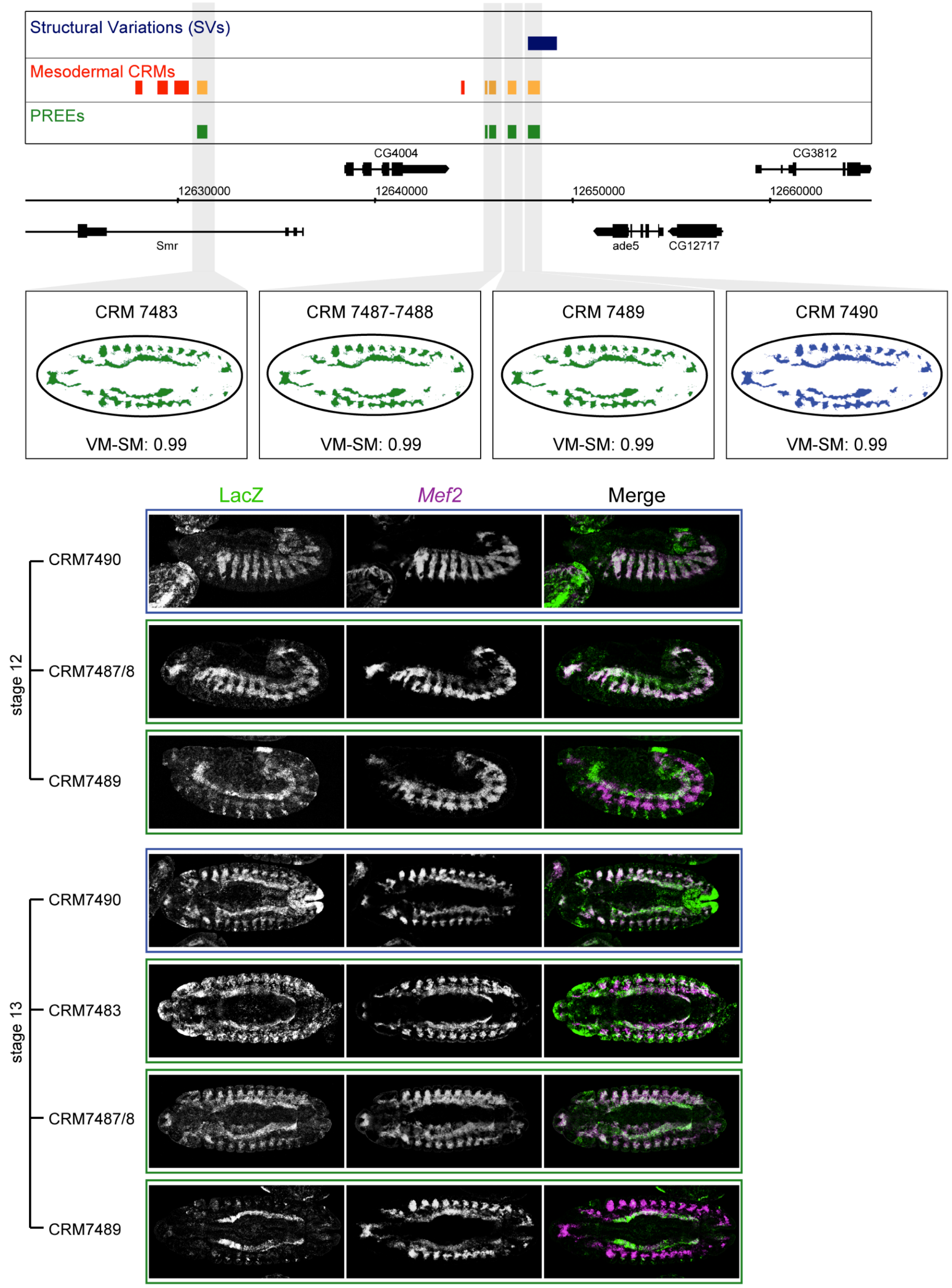


Fig. 9 PREEs' spatio-temporal activity overlap in the ade5 locus.

4 PREEs were predicted in the locus with similar in the predicted activity (green). One of the PREEs (CRM7490) is completely deleted by structural variations in one isogenic line (blue). SVM specificity scores and predicted expressions of the CRMs tested in a reporter assay are shown. Double fluorescent ISHs of the lacZ reporter (green) driven by PREEs (CRM 7490,

7483, 7487/88, 7489) and a general mesodermal marker, Mef2 (magenta), show a similar activity of the PREEs (compare LacZ expression). In particular CRM7490, CRM7487/8 (stage 12 and 13) and CRM 7483 (stage 13) show activity in somatic and visceral mesoderm while CRM 7489 is active only in the visceral mesoderm (stage 12 and 13). All embryos are oriented anterior - left, dorsal – up.

This complex locus demonstrates partial redundancy at both spatial and temporal level, where many enhancers with overlapping expression are likely required to ensure robust expression.

Different is the case of the *Traf4* locus (Fig. 10). *Traf4* is a member of the tumor necrosis factor receptor superfamily and induces the activation of the c-Jun N-terminal kinase pathway. *Traf4* Null mutants failed to develop into the pupal stage because of the gene role in the development of imaginal eye discs and the formation of a correct photosensory neuronal array in the brain hemisphere (Cha et al. 2003; WAGNER 2008). We found that three putative redundant enhancers (CRMs 5429, 5432, 5435/6) had share territory of activity at embryonic stage 6 in line with their prediction. However, we found that two other enhancers (CRMs 5437, 5440) which were not predicted redundant, because just below the threshold that we used to define putative redundant enhancers, are active also early in the mesoderm at stage 6. Therefore, although the total spatial expression pattern of each enhancer varies, they are all active in a population of mesodermal cells at the same stage of development. These results highlight the complexity of Traf4 regulation and the extent to which one enhancer activity may be buffered by additional elements at a given developmental stage.

A similar complex case is the *Caderin-N* (*CadN*) locus. *CadN* is important for the cell-cell interaction during gastrulation and in the embryonic nervous system (Iwai et al. 1997; Alon 2007; Oda et al. 1998; Alon et al. 1999). The dynamic interplay between CadN and shotgun in early stages of *D. melanogaster* development was proved to be important for the epithelial-mesenchymal transition (Oda et al. 1998; Becskei and Serrano 2000).

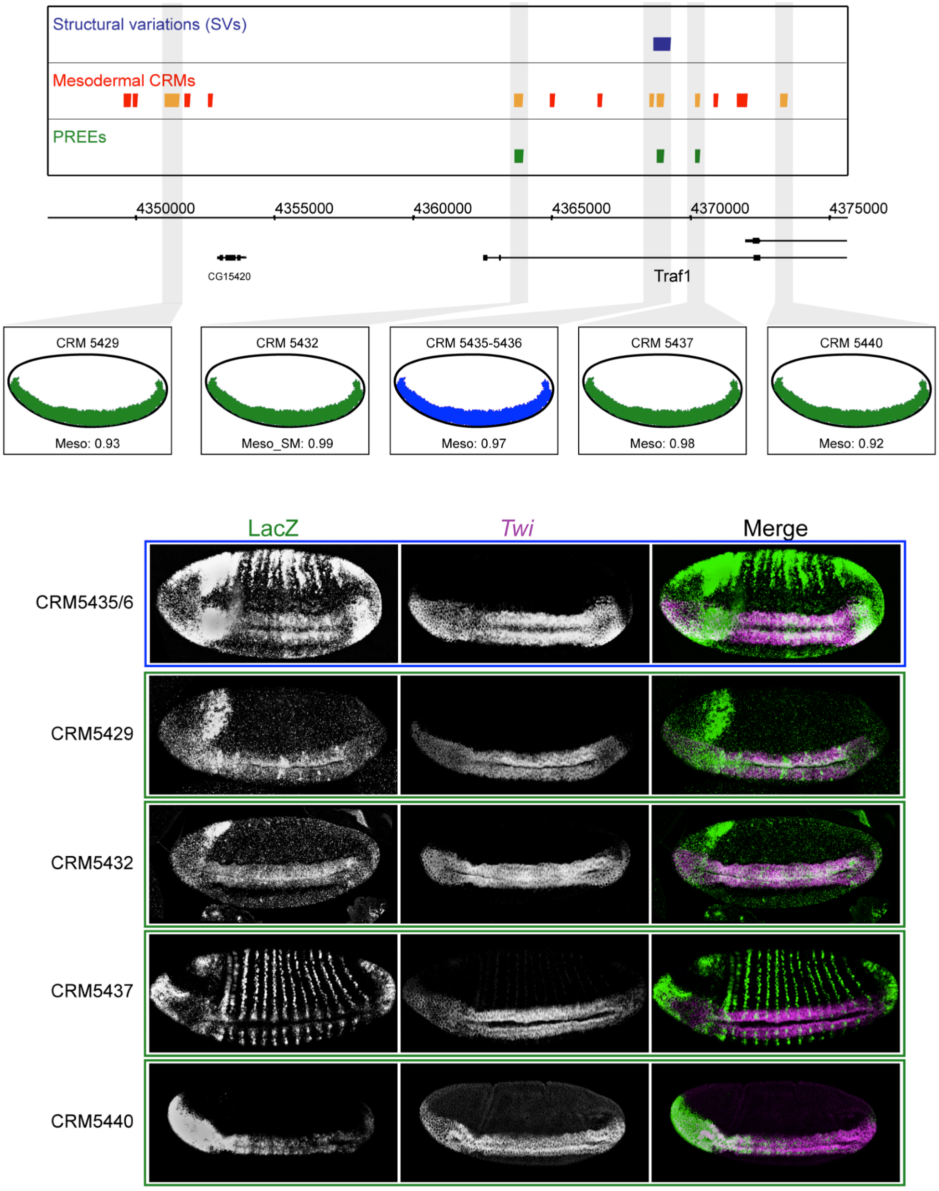


Fig. 10 Complexity of Traf4 regulation.

The Traf4 locus is characterized by the presence of 3 PREEs (green) predicted because of their similar activity and where the CRM 5435/6 is deleted (blue). Since some other CRMs in the region have a SVM specificity score just below the threshold (that is >0.95) we analyzed their activity in a reporter assay as well. Double fluorescent ISHs (bottom) was performed using probes for LacZ (green) and a marker expressed early in the mesoderm Twist (magenta). All five CRMs have activity early in the mesoderm (stage 6) that partially overlaps with each other. All embryos are oriented anterior - left, dorsal – up.

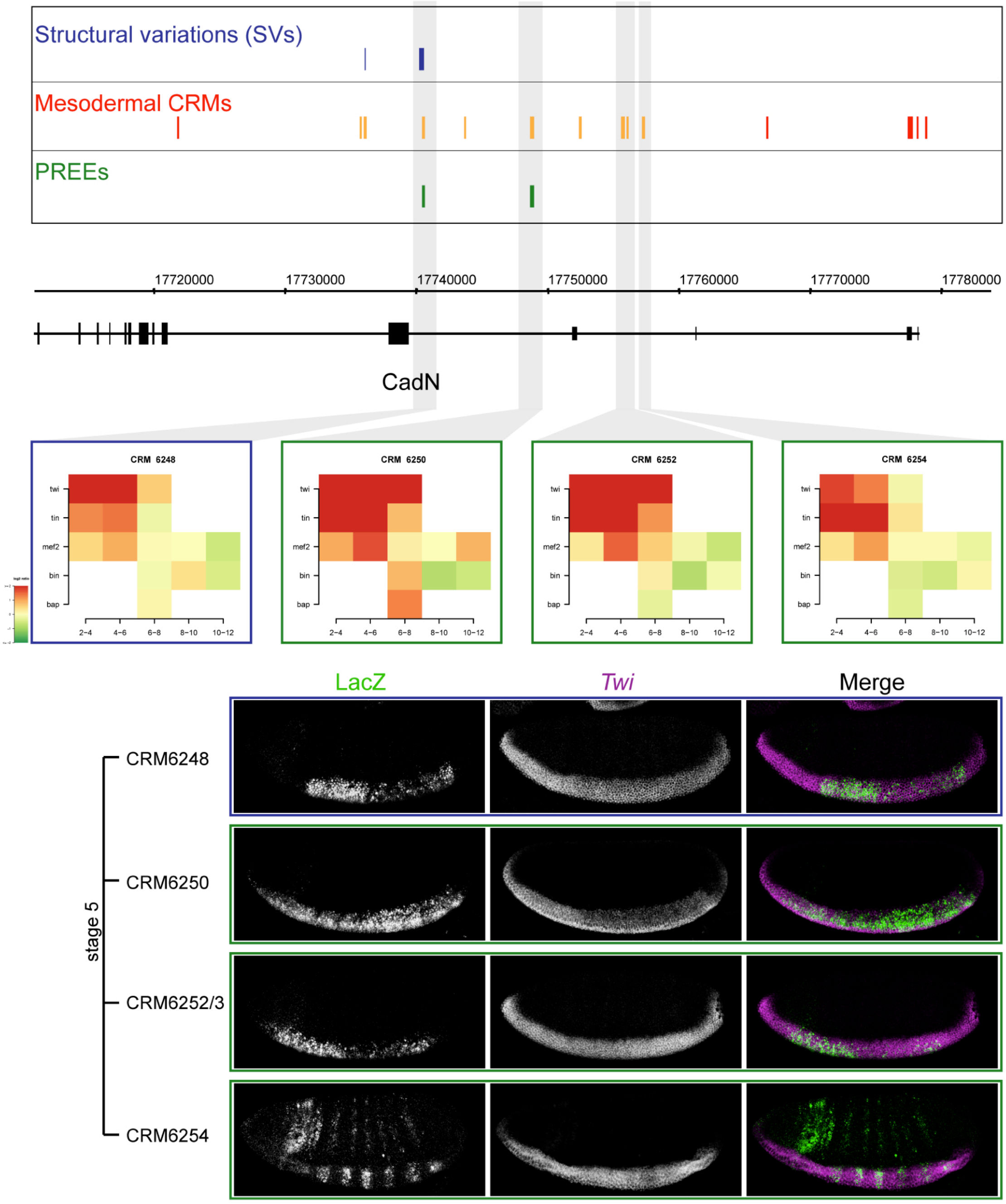


Fig. 11 Complexity of Traf4 regulation.

The Traf4 locus is characterized by the presence of 3 PREEs (green) predicted because of their similar activity and where the CRM 5435/6 is deleted (blue). Since some other CRMs in the region have a SVM specificity score just below the threshold (that is >0.95) we analyzed their activity in a reporter assay as well. Double fluorescent ISHs (bottom) was performed using probes for LacZ (green) and a marker expressed early in the mesoderm Twist (magenta). All five CRMs have activity early in the mesoderm (stage 6) that partially overlaps with each other. All embryos are oriented anterior - left, dorsal – up.

In this locus, we predicted two putative redundant enhancers with similar transcription factor binding profile (Fig. 11) and one of PREEs deleted in two isogenic lines. Anyway we investigated the activity of five more enhancers since they had similar binding signatures just below our Spearman cut off. Examination of the *lacZ* expression revealed that 4 of these enhancers have overlapping expression at stage 5 (Fig. 11).

Taken together, these data indicate that partial redundant enhancers are pervasive throughout the genome. Although this study provides the first systematic attempt to estimate how frequently redundancy in transcriptional regulation occurs, our predictions clearly underestimate the number of enhancers with partially overlapping activity within a given gene's locus. Moreover we have found that the level of potential redundancy is much more complex than previously observed. In approximately 50% of the cases not simply two enhancers may act redundantly, but there are three (rols), four (CadN and ade5) or even five (Traf1) enhancers with overlapping spatio-temporal activity. This extensive level of potential *cis*-regulatory redundancy is not just present in the loci of TFs, which are a class of proteins known to have complex transcriptional regulation and many enhancer elements (Lenhard et al. 2012; Alon et al. 1999; Alon 2007), but it's rather present in a wide range of genes. Here, we purposely chose gene loci of proteins with diverse functions; CadN is an adhesion protein, Traf1 a signaling receptor, Rols is an intracellular adaptor protein, ade5 is a metabolic protein and CG42788 is a predicted component of the E3 ubiquitin-protein ligase complex.

Conclusion

For decades the presence of redundant elements in the regulation of gene expression has been the object of a discussion that recently has been rekindled. Many works have concentrated on the regulation of a single gene by two redundant elements identified using different criteria (Perry et al. 2009; Rosenfeld et al. 2002; Perry et al. 2011; 2012; Montavon et al. 2011; Cretekos et al. 2008; Degenhardt et al. 2010; Xiong et al. 2002). Only recently it was observed that several partial redundant enhancer could work together to ensure the robustness of the expression level (Frankel et al. 2010; Becskei and Serrano 2000). Moreover it was estimated that 50% of the target genes of the transcription factor Dorsal contain redundant enhancers, suggesting that the presence of additional enhancer may be typical for key developmental genes (Perry et al. 2009; Becskei and Serrano 2000; Frankel et al. 2010). Here we present a systematic study of the role of PREEs across the genome in a developmental context. Our results reveal a more complex transcriptional regulation at least in a developmental contest. PREEs appear to be a pervasive part of the regulation genes expressed during embryonic development. We were able to identify 1942 that are likely acting redundantly with another element in the sense of driving similar patterns of expression and having the same predicted target gene. Looking globally, our predictions identified a partially redundant pair in 71% of gene loci in which we predicted PREEs. However the other 29% of loci contained three, four, five or even more elements predicted to have partially redundant activity. We specifically selected some of these complex loci for in vivo transgenic reporter analysis, and found that 50% loci that we investigated in detail had at least three (in one case even five) partially redundant enhancers.

However our extensive in vivo analysis indicates that this is almost certainly an underestimate of the extent of redundancy: when we tested the activity of regions just below the cut-offs we used to define PPREs. This unanticipated level of redundancy changes the classic paradigm of redundancy between two elements suggesting that partially redundant enhancers are pervasive throughout the genome and therefore could have a substantial impact on the robustness of gene expression during embryonic development.

In agreement with the strict definition of redundancy, we observed that deletion of a redundant enhancer does not cause major phenotypic alteration in a wid-type population, at least in a given environmental condition. What then prevents the deletion of PPREs in an organism?

The answer may partially lie in the fact that these elements could play complementary roles in fluctuating environments. In stable and optimal conditions, deletions of one of the multiple PREEs are tolerated. This can be explained by the presence of alternative elements that could compensate for this lost due to the partial overlapping activity. This is the case of svb where the deletion of one of the partially redundant enhancers do not cause major phenotypic variations (Frankel et al. 2010; Alon et al. 1999). In our study we also observe that deletion of enhancers are allowed when flies are grown in a controlled environment. It is possible though that in the wild where organisms face constantly environmental variations, the specific contribution of the deleted enhancers to the expression of the target gene is necessary for the fitness. Having then multiple enhancers with partial overlapping activity ensures the correct level of expression of the gene, thus conferring genetic and environmental robustness.

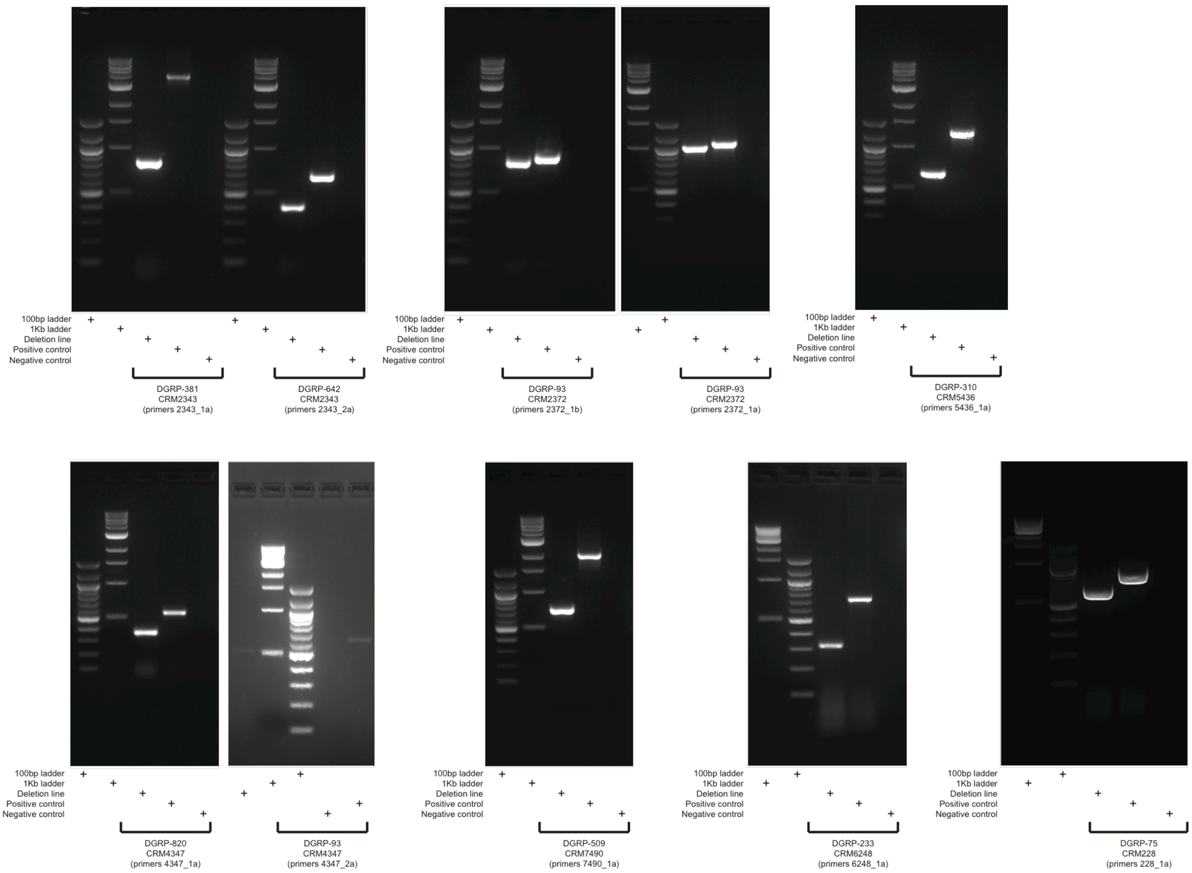
In another scenario, these elements often drive overlapping patterns of expression, there are at least some tissues, stages, or environmental conditions in which the elements play distinct functional roles. The overlap in activity can be restricted to a small time window or a small number of cells, while in others the overlap is more extensive in time or space (Fig. 9 and 10). Thus, while an enhancer may be redundant with another element in one tissue or developmental stage, its activity may be non-redundant in another cell type and therefore be essential for embryonic development. Similarly, enhancers that appear redundant in 'normal' environmental conditions could act non-redundantly when the environmental conditions become more extreme, as recently observed in the svb locus (Frankel et al. 2010; Alon et al. 1999). It is this partially redundant property that most likely holds the key to how these elements are maintained.

A previous study hypothesized that there may be different evolutionary pressure on two redundant enhancers: the primary enhancer being more constrained then the redundant shadow enhancer, allowing the later to accumulate mutations without inducing a phenotype and thus evolve faster (Hong et al. 2008; Alon et al. 1999). However, our analyses of sequence conservation and the frequency of segregating mutations affecting these enhancers suggests that the evolutionary pressures affecting these are similar, and overall show a stronger tendency towards conservation than do more isolated elements driving similar expression profiles. This strongly suggests that enhancers with overlapping activity are being maintained for a purpose. One such property of many PREEs is their non-redundant activity, which may be under selective pressure while their redundancy is maintained as a beneficially consequence. Alternatively, ‘redundant’ enhancers driving similar spatio-temporal activity could act together to guarantee that a gene reaches a certain level of expression or could have essential roles in ensuring correct patterning precision (Perry et al. 2011; Barkai and Leibler 1997; Perry et al. 2010; Yi et al. 2000; Dunipace et al. 2011) or reduce stochastic effects on gene expression (Perry et al. 2011; Maeda and Sano 2006). In these cases PREEs ensure robustness of the trait when environmental variations occur but do not confer genetic robustness to all possible mutations since for example deletions of a partially redundant enhancer drastically influence the viability of an organism (Dunipace et al. 2011; Alon 2007).

We therefore argue that redundant enhancers have likely evolved to provide robustness to gene expression in the context of fluctuating environmental and genetic perturbations. Indeed, redundant enhancers are important to maintain correct gene expression in different environmental conditions (Frankel et al. 2010; Dassow et al. 2000; Perry et al. 2010; Dunipace et al. 2011). In addition, our results indicate that PREEs facilitate normal development of a natural wild type population to occur despite numerous structural variations affecting many regulatory elements. Therefore the presence of PREEs also confers genetic robustness during embryonic development.

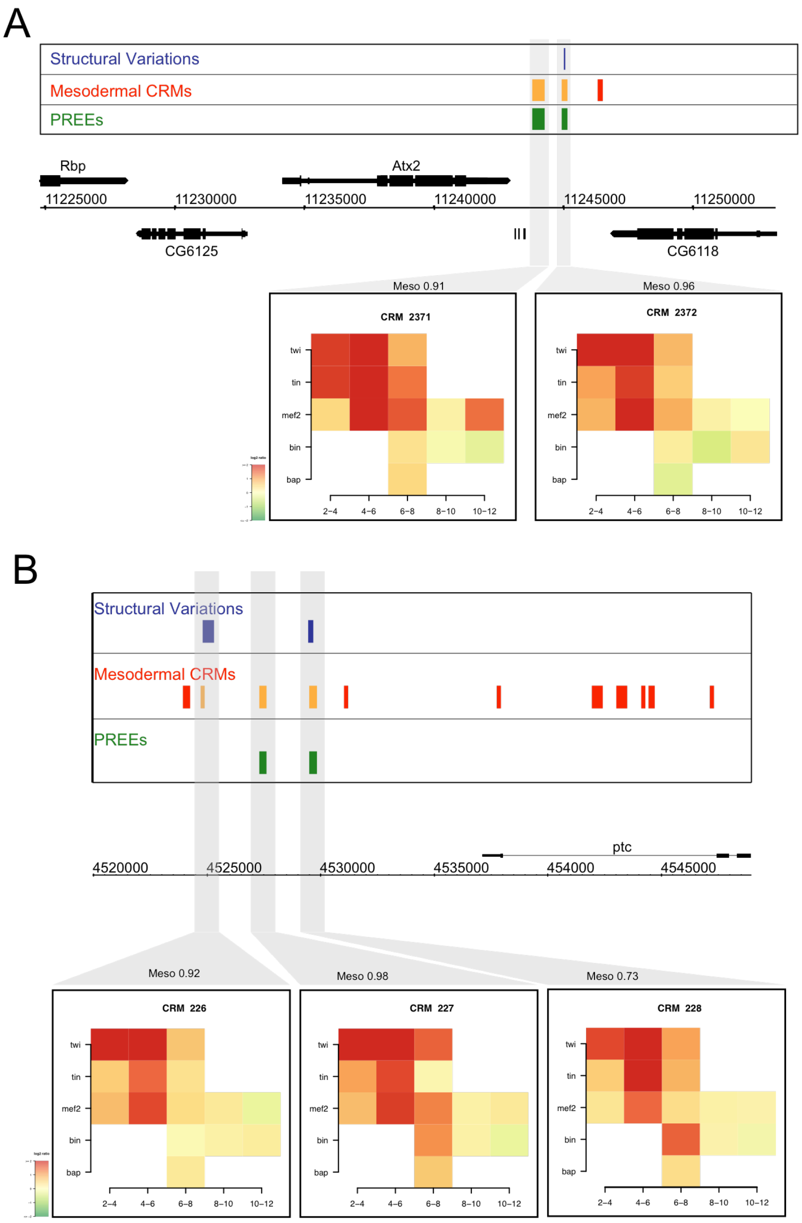
Therefore we have shown that multiple partially redundant enhancers appear to be a fundamental component of transcriptional regulation during embryonic development, where they can confer phenotypic robustness through a diverse range of mechanism. The extensive nature of these elements overlapping activity also adds another layer of complexity to the flow of information through large developmental networks, a role that has yet to be explored. Despite their partial overlap, PREEs have distinct functions at least in some cell or tissues or in some environmental so that it results in selection maintaining them throughout evolution. This doesn't mean, of course, that the overlapping nature of enhancers does not provide opportunities for evolutionary innovation: however it indicates that these elements are playing partially independent functional roles in development.

Supplementary Figures and Tables



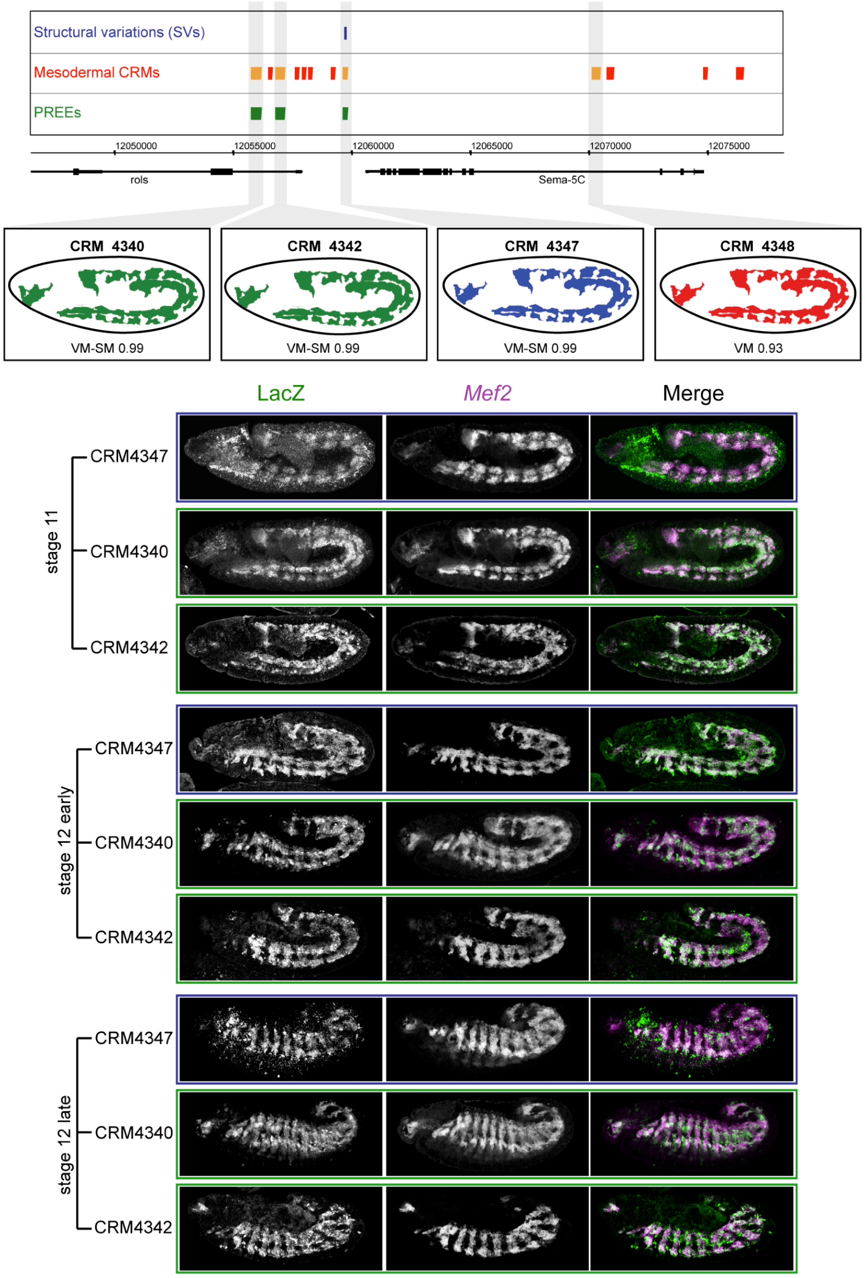
Suppl. Fig. 1 Validation of enhancer deletion by PCR.

We used the indicate primers (Suppl. table 2) to validate the predicted deletion of enhancers in one or two DGRP lines. DNA extracted from the reference line 2057 (Bloomington Stock Number) was used as positive control.



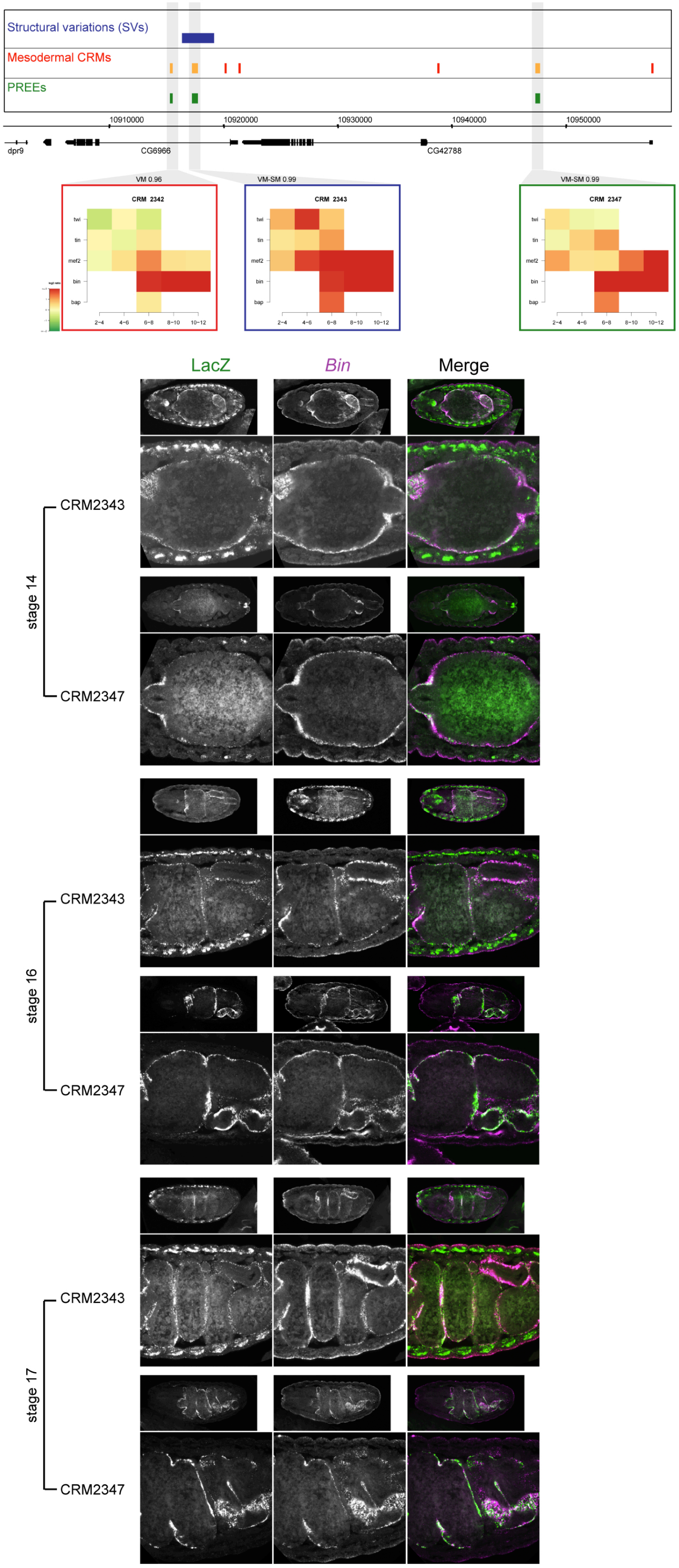
Suppl. Fig. 2 Unobserved partially redundancy in Atx2 and ptc loci.

We analyzed the activity of predicted PREEs (green) in the two loci depending on their TF occupancy profile similarity (heatmaps bottom). One of the PREEs is deleted (blue) in some isogenic lines a specific line. We observed that one of the PREEs did not have any activity in a transgenic reporter assay.



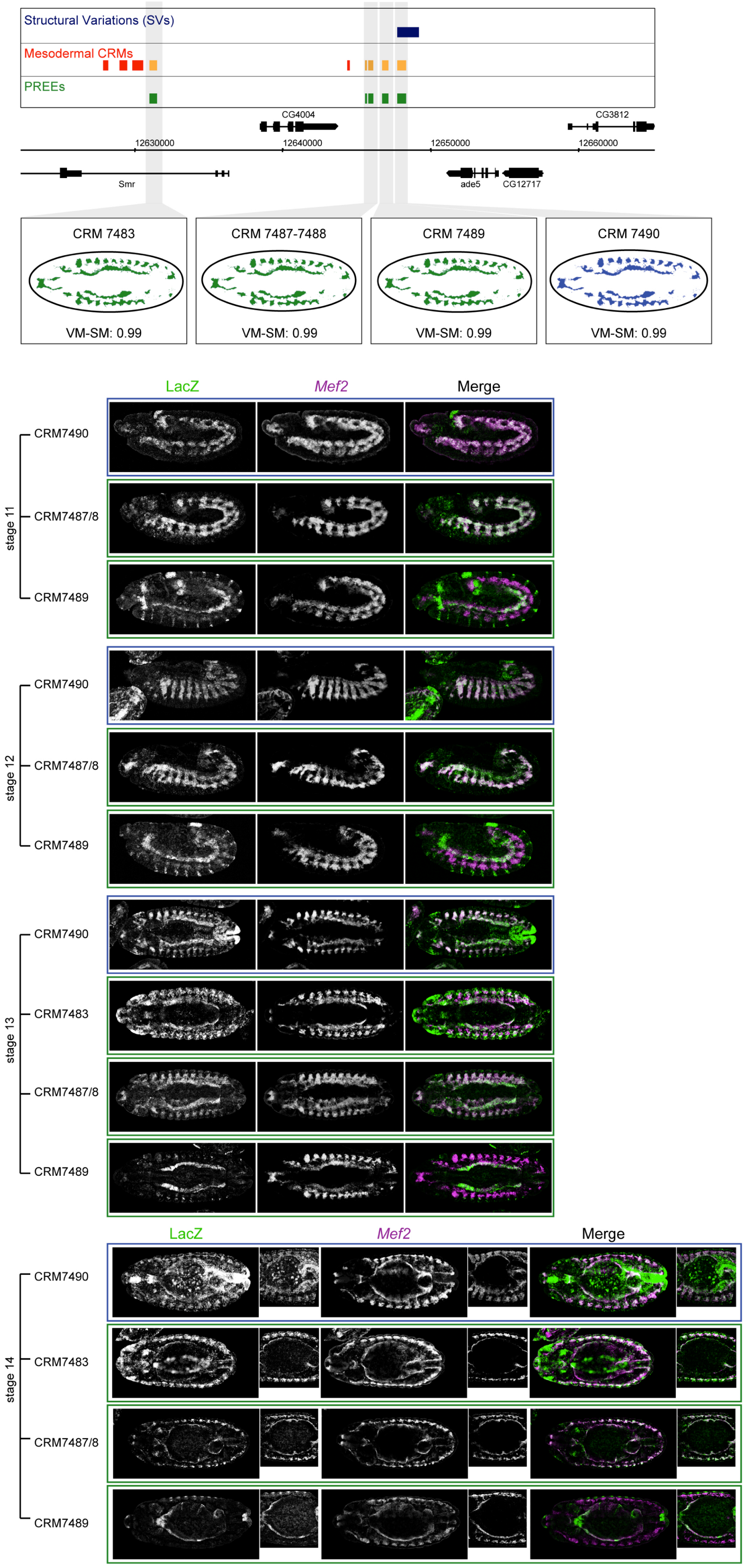
Suppl. Fig.3 *rols* locus.

CRM4347 is affected by deletion in several DGRP lines and predicted to be partially redundant with the CRM4340 and CRM4342 (green). CRM4348 (red) was included in the reporter assay since the SVM specificity score is slightly below the threshold (>0.95). However we did not observe overlapping activity with the deleted enhancer. Double fluorescent ISHs (bottom) of the lacZ reporter (green) driven by PREEs (CRM4347, 4340, 4342) and a general mesodermal marker, Mef2 (magenta), show a similar activity in the somatic and visceral mesoderm of the PREEs from embryonic stage 11 to late stage 12 (compare LacZ expression).


Structural variations (SVs)
Mesodermal CRMs
PREEs
10910000
10920000
10930000
10940000
10950000
dpr9
CG6966
CG42788
VM 0.96
VM-SM 0.99
VM-SM 0.99
CRM 2342
CRM 2343
CRM 2347
twi
tin
mef2
bin
bap
2-4
4-6
6-8
8-10
10-12
LacZ
Bin
Merge
stage 14
CRM2343
CRM2347
stage 16
CRM2343
CRM2347
stage 17
CRM2343
CRM2347

Suppl. Fig.4 CG42788 locus.

We identified 3 PREEs with high similar TF occupancy profile (heatmap) and similar predicted activity (SVM prediction and specificity score). CRM2343 is completely deleted in some lines from the DGRP panel. Despite the CRM2342 was predicted to be a PPRE, it did not share regions of overlap with the other PREEs. Double fluorescent ISHs (bottom) of the lacZ reporter (green) driven by PREEs (CRM2343, CRM2347) and a general mesodermal marker, Mef2 (magenta), show a similar activity of the PREEs in visceral mesoderm from stage 14 to 17 (compare LacZ expression). However CRM2343 shows non-overlapping activity in the somatic mesoderm.


Structural Variations (SVs)
Mesodermal CRMs
PREEs
CG4004
CG3812
12630000
12640000
12650000
12660000
Smr
ade5
CG12717
CRM 7483
VM-SM: 0.99
CRM 7487-7488
VM-SM: 0.99
CRM 7489
VM-SM: 0.99
CRM 7490
VM-SM: 0.99
LacZ
Mef2
Merge
stage 11
CRM7490
CRM7487/8
CRM7489
stage 12
CRM7490
CRM7487/8
CRM7489
stage 13
CRM7490
CRM7483
CRM7487/8
CRM7489
LacZ
Mef2
Merge
stage 14
CRM7490
CRM7483
CRM7487/8
CRM7489

Suppl. Fig.5 ade5 locus.

4 PREEs with similar in the predicted activity were identified (green) (SVM prediction and specificity score are shown). One of the PREEs (CRM7490) is completely deleted by structural variations in one isogenic line (blue). Double fluorescent ISHs of the lacZ reporter (green) driven by PREEs (CRM 7490, 7483, 7487/88, 7489) and a general mesodermal marker, Mef2 (magenta), show a similar activity of the PREEs (compare LacZ expression). In particular CRM7490, CRM7487/8 (from stage 11 to stage 14) and CRM 7483 (stage 13 and 14) show activity in somatic and visceral mesoderm while CRM 7489 is active only in the visceral mesoderm (from stage 11 to stage 14). This partial and temporal overlap between the PREEs is an indication of the redundancy complex role to ensure robustness to the phenotype. All embryos are oriented anterior - left, dorsal – up

Suppl. Table 2

Primer_set_id	Forward_primer	Reverse_primer
228_1a	CAGCCAATGTTCTCGGAAGGTA	CGAAATGCGGGCTGAAGTTT
2343_1a	TCGTGCTCCTACAACAGGCTAA	GTGCATACGCGAATACGTGAC
2343_2a	ACTTCGTGTACGGCGTTCAACT	GTTTATAAGGTCGCATTGCCAG
2372_1a	CTTGCTAGTAAGGATCTCCGGT	ACCCTGACAGGCTGGGAGAC
2372_1b	TTCGAGACTTTTAGGGGAAGGACTG	ATACACACTACAAGCGCCGACTAGC
4347_1a	CAAGATGCATCCGTGTACTAAG	TCGTAGTTCCATAAACGAACAGAA
4347_2a	CAAGATGCATCCGTGTACTAAG	TTAGTCCACGTCAGCCAAAA
5436_1a	CCCAGTTTACACTCTAATGAGGGA	TTTTACGAGGAAGGAAGGCGT
6248_1a	TTCACTCAGAAAGCAGGttattaaa	GCCATTGTCGGATTAAGGGA
7490_1a	TTTTTGGTACACATGACAGAGTCG	CGATCGCTCTACACGTTTTCAT
7490_2a	TTGGTACACATGACAGAGTCGAAA	tatatatCGGGAATTGGCAGGG
7490_2b	CTTATCACTTGTCCTACCTCCCGCT	GTTGCTGTAACCGTCGAttgttgtt

Primer_set_id	SV_size	Primer_dist	Primer_dist_SV	Chr	Pos_FP	Pos_RP
228_1a	249	899	649	chr2R	4529091	4529989
2343_1a	2856	3671	814	chr3R	10915794	10919464
2343_2a	257	649	391	chr3R	10917517	10918165
2372_1a	64	1093	1028	chr3R	11244543	11245635
2372_1b	64	871	806	chr3L	12059414	12059968
4347_1a	184	555	370	chr3L	12059414	12060023
4347_2a	101	610	508	chr2L	4368202	4369499
5436_1a	648	1298	649	chr2L	17740006	17740716
6248_1a	391	711	319	chrX	12647396	12649602
7490_1a	1477	2207	729	chrX	12647399	12648741
7490_2a	121	1343	1221	chr3R	11244504	11245374
7490_2b	121	500	378	chrX	12647974	12648473

Suppl. Table 3

Seq ID	Seq FORWARD	Seq REVERSE
227|227|2R|4527273|4527701	GGTACCGATGGGTGTGGGGTAAATCC	AGATCTCACGAATTAACCTACCTACAAATTACG
228|228|2R|4529475|4529921	AGATCTGAATTCCATAAAAAGTGAAAGATTCC	GGTACCGAAAAGAAGAAACGCAAGAGC
2343|2343|3R|10917226|10917889	GGCGCGCCGGGTTTCTGCAGCTTTTAAGG	GGTACCGTATACGCCGAAGAGTGTGC
2372|2372|3R|11244909|11245262	GGTACCCCTTGGCTCTTTCACCTTCC	AGATCTTGGAAATGGAGAAGGAATAGG
4340|4340|3L|12055730|12056322	AGATCTAGCAGTTCCCACATGGAAGG	GGTACCGCAACAATAACTATTGCAACACC
5438|5438|2L|4370814|4371114	GGTACCGTTGTTGCCGGTTTTTATGC	AGATCTCAAACGCAGCTCTAATGACG
5440|5440|2L|4373198|4373596	GGTACCTGAGAGCAATTTGCCAAAGC	AGATCTCACCTTATTGTTAGGTAATCGATGG
6246+6247|6246+6247|2L|17735713|17736332	AGATCTTTTCATTTGTTTCAGGTTTTGG	GGTACCTTCTTTCCCAACAAATATCAGG
6248|6248|2L|17740445|17740788	GGTACCTCACAGTCTGTTCCAGTCAAGG	AGATCTGAACCTTTTTCCGACTTACCC
6250|6250|2L|17748638|17749061	GGTACCGCTTGCATTTAAACAAATTTTCC	AGATCTGCTCTTCTTGCACTTATTTTTGC
6251|6251|2L|17752418|17752790	GGTACCTGTGTGTCTCTGTATATCCTTTGG	AGATCTCACTGAGAGTTTTCCTTGTGG
6254|6254|2L|17757163|17757496	GGTACCGTAATTTTAAGACAACAAACACAAGC	AGATCTTTCCTATAAACAACTAATGAGAGTGG
7483|7483|X|12630965|12631616	GGCGCGCCTGCATTCGTTACTACATTGTTCC	GGTACCTGCATGAATAAAACAGTCAAAAGC
7487+7488|7487+7488|X|12645540|12646228	GGTACCTTGGTTGCTTCTTCTTCTTGC	GGCGCGCCTCTTCATCATCATCATCTTATTCC
7489|7489|X|12646711|12647253	GGTACCCTTGTGGCAAACAAACTTGC	AGATCTGCGTTGGAATCTCTGTCTCC
226|226|2R|4524627|4525027	GGTACCACGAACGAACAACCAAAACG	AGATCTGATCCTTCTCCCCATTCTCG
2342|2342|3R|10915256|10915707	AGATCTAAAACGCTCCACGATAAAGC	GGTACCCCATCTTTCTGTTTATCACGATCC
2347|2347|3R|10947246|10947903	AGATCTAATATTGAATGGAGAGTCCTTGG	GGTACCACAAATCCACTTGGCAGAGC
4347|4347|3L|12059533|12059978	AGATCTTGTTTCAATTTCTCGCTTTCC	GGTACCCAACCAATGGTCGTAGTTCC
4348|4348|3L|12070039|12070669	AGATCTAGTTTGGGTGGTTCCTAGCC	GGTACCGCTCGAATTAATGTTCACTGG
5429|5429|2L|4350975|4351728	AGATCTTTTTAACCCGCAGCATTTCC	GGTACCTGGCAATTAATCACGCAAGC
5432|5432|2L|4363552|4364106	GGTACCCATAATAAAATCTTTTGGCAAGC	AGATCTAAGCATCGACCTGATCTTCC
5435+5436|5435+5436|2L|4368442|4369166	GGTACCGAAGTGCAGCAAATGAGTCG	GGCGCGCCCCTACTGCTGCTGATGTTCG
6249|6249|2L|17743625|17744025	GGTACCCATTACGCACTCGTCTCTGG	AGATCTTCTTGAGGTCTGTGGTCTGG
6252+6253|6252+6253|2L|17755542|17756384	GGTACCGAATTGATTGGATTGCATCG	AGATCTGTCCTTTGGTGACTAATGAAGC
7490|7490|X|12647697|12648516	AGATCTACATATGGTAGATAAGATGGAAACG	GGTACCGCTCACTGTACCTCATTGTTGC
2371|2371|3R|11243660|11244448	GGTACCACGACGTAAGGCGTTTAGGG	AGATCTCTGGGAAGGGCATGTTAAGG
4342|4342|3L|12056657|12057408	AGATCTATGCCTTCATTCTTCTGTCG	GGTACCGTCCTAAAAAGGGAGCATCG

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–984.

Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T, Thanos D. 2000. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103**: 667–678.

ALATALO RV, GUSTAFSSON L, LINDEN M, LUNDBERG A. 1985. Interspecific Competition and Niche Shifts in Tits and the Goldcrest: An Experiment. *Journal of Animal Ecology* **54**: 977–984.

Alatalo RV, Moreno J. 1987. Body Size, Interspecific Interactions, and Use of Foraging Sites in Tits (Paridae). *Ecology* **68**: 1773–1777.

Alon U. 2007. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**: 450–461.

Alon U, Surette MG, Barkai N, Leibler S. 1999. Robustness in bacterial chemotaxis. *Nature* **397**: 168–171.

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. 2009. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**: 47–57.

Arnosti DN, Barolo S, Levine M, Small S. 1996a. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**: 205–214.

Arnosti DN, Barolo S, Levine M, Small S. 1996b. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**: 205–214.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Banerji J, Rusconi R, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27:** 299-308

Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. *Nature* **387**: 913–917.

Becskei A, Serrano L. 2000. Engineering stability in gene networks by autoregulation. *Nature* **405**: 590–593.

Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, Miranda TB, Sung M-H, Trump S, Lightman SL, et al. 2011. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* **43**: 145–155.

Brown CD, Johnson DS, Sidow A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**: 1557–1560.

Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet*.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626–635.

Castanon I, Stetina Von S, Kass J, Baylies MK. 2001. Dimerization partners determine the activity of the Twist bHLH protein during Drosophila mesoderm development. *Development* **128**: 3145–3159.

Cha G-H, Cho KS, Lee JH, Kim M, Kim E, Park J, Lee SB, Chung J. 2003. Discrete functions of TRAF1 and TRAF2 in Drosophila melanogaster mediated by c-Jun N-terminal kinase and NF-kappaB-dependent signaling pathways. *Mol Cell Biol* **23**: 7982–7991.

Chapman LJ, Galis F, Shinn J. 2000. Phenotypic plasticity and the possible role of genetic assimilation: Hypoxia-induced trade-offs in the morphological traits of an African cichlid. *Ecology Letters* **3**: 387–393.

Chen K, van Nimwegen E, Rajewsky N, Siegal ML. 2010. Correlating Gene Expression Variation with cis-Regulatory Polymorphism in Saccharomyces cerevisiae. *Genome Biol Evol* **2**: 697–707.

Clarke G, McKenzie J. 1987. Developmental stability of insecticide resistant phenotypes in blowfly; a result of canalizing natural selection. *naturecom*.

Cooke J, Nowak MA, Boerlijst M, Maynard-Smith J. 1997. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet* **13**: 360–364.

Cretekos CJ, Wang Y, Green ED, Martin JF, Rasweiler JJ, Behringer RR. 2008. Regulatory divergence modifies limb length between mammals. *Genes Dev* **22**: 141–151.

Dassow von G, Meir E, Munro EM, Odell GM. 2000. The segment polarity network is a robust developmental module. *Nature* **406**: 188–192.

Day T, Pritchard J, Schluter D. 1994. A comparison of two sticklebacks. *EVOLUTION* 1723–1734.

De Jong G. 1995. Phenotypic Plasticity as a Product of Selection in a Variable Environment. *American Naturalist* **145**: 493–512.

de Jong G. 2005. Evolution of phenotypic plasticity: patterns of plasticity and the emergence of ecotypes. *New Phytol* **166**: 101–117.

Degenhardt KR, Milewski RC, Padmanabhan A, Miller M, Singh MK, Lang D, Engleka KA, Wu M, Li J, Zhou D, et al. 2010. Distinct enhancers at the Pax3 locus can function redundantly to regulate neural tube and neural crest expressions. *Dev Biol* **339**: 519–527.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**: 390–403.

Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA. 2012. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**: 1233–1244.

Denger JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK. 2012 DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390–394

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

DUN RB, FRASER AS. 1958. Selection for an invariant character; vibrissa number in the house mouse. *Nature* **181**: 1018–1019.

Dunipace L, Ozdemir A, Stathopoulos A. 2011. Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. *Development* **138**: 4075–4084.

Dunipace L, Saunders A, Ashe HL, Stathopoulos A. 2013. Autoregulatory Feedback Controls Sequential Action of cis-Regulatory Modules at the brinker Locus. *Dev Cell* **26**: 536–543.

Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res* **17**: 1898–1908.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.

Eshel I, Matessi C. 1998. Canalization, genetic assimilation and preadaptation. A quantitative genetic model. *Genetics* **149**: 2119–2133.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**: 610–618.

FEINSINGER P, SWARM LA. 1982. “Ecological Release,” Seasonal Variation in Food Supply, and the Hummingbird Amazilia Tobaci on Trinidad and Tobago. *Ecology* **63**: 1574–1587.

Flatt T. 2005. The evolutionary genetics of canalization. *Q Rev Biol* **80**: 287–316.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490–493.

Frankel N, Erezyilmaz DF, Mcgregor AP, Wang S, Payre F, Stern DL. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* **474**: 598–603.

Frasch M. 1995. Induction of visceral and cardiac mesoderm by ectodermal Dpp in the early Drosophila embryo. *Nature* **374**: 464–467.

Futuyma DJ. 2009. Evolution. Sunderland, Mass.: Sinauer Associates

Ghiasvand NM, Rudolph DD, Mashayekhi M, Brzezinski JA, Goldman D, Glaser T. 2011. Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. *Nat Neurosci* **14**: 578–586.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63–66.

Gurevitch J. 1992. Sources of variation in leaf shape among two populations of Achillea lanulosa. *Genetics* **130**: 385–394.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**: 269–276.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* **402**: C47–52.

Higashijima S, Michiue T, Emori Y, Saigo K. 1992. Subtype determination of Drosophila embryonic external sensory organs by redundant homeo box genes BarH1 and BarH2. *Genes Dev* **6**: 1005–1018.

HOGSTAD O. 1978. DIFFERENTIATION OF FORAGING NICHE AMONG TITS, PARUS SPP., IN NORWAY DURING WINTER. *Ibis* **120**: 139–146.

Hong J-W, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**: 182–192.

Ip YT, Kraut R, Levine M, Rushlow CA. 1991. The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in Drosophila. *Cell* **64**: 439–446.

Iwai Y, Usui T, Hirano S, Steward R, Takeichi M. 1997. Axon Patterning Requires DN-cadherin, a Novel Neuronal Adhesion Receptor, in the Drosophila Embryonic CNS. *Neuron* **19**: 77–89.

Jakobsen JS, Braun M, Astorga J, Gustafson EH, Sandmann T, Karzynski M, Carlsson P, Furlong EEM. 2007. Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes Dev* **21**: 2448–2460.

Jaynes JB, Fujioka M. 2004. Drawing lines in the sand: even skipped et al. and parasegment boundaries. *Dev Biol* **269**: 609–622.

Jaźwińska A, Kirov N, Wieschaus E, Roth S, Rushlow C. 1999. The Drosophila gene brinker reveals a novel mechanism of Dpp target gene regulation. *Cell* **96**: 563–573.

Jongens TA, Fowler T, Shermoen AW, Beckendorf SK. 1988. Functional redundancy in the tissue-specific enhancer of the Drosophila Sgs-4 gene. *EMBO J* **7**: 2559–2567.

Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EEM. 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**: 473–486.

Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nat Genet* **37**: 295–299.

Keys DN, Lewis DL, Selegue JE, Pearson BJ, Goodrich LV, Johnson RL, Gates J, Scott MP, Carroll SB. 1999. Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* **283**: 532–534.

Kitano H. 2004. Biological robustness. *Nat Rev Genet* **5**: 826–837.

Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics* **76**: 8–32.

Korge G. 1975. Chromosome puff activity and protein synthesis in larval salivary glands of Drosophila melanogaster. *Proc Natl Acad Sci USA* **72**: 4550–4554.

Krakauer DC, Plotkin JB. 2002. Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci USA* **99**: 1405–1409.

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130**: 6569–6575.

Langkjaer RB, Cliften PF, Johnston M, Piskur J. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421**: 848–852.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**: 233–245. http://www.nature.com/nrg/journal/v13/n4/full/nrg3163.html.

Leptin M. 1991. twist and snail as positive and negative regulators during Drosophila mesoderm development. *Genes Dev* **5**: 1568–1576.

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**: R754–63.

Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. 2009. A microRNA imparts robustness against environmental fluctuation during development. *Cell* **137**: 273–282.

Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. 2011. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol* **12**: R34.

Liao B-Y, Zhang J. 2007. Mouse duplicate genes are as essential as singletons. *Trends Genet* **23**: 378–381.

Little SC, Tikhonov M, Gregor T. 2013. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* **154**: 789–800.

Liu Y-H, Jakobsen JS, Valentin G, Amarantos I, Gilmour DT, Furlong EEM. 2009. A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Dev Cell* **16**: 280–291.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

Loots GG, Kneissel M, Keller H, Baptist M, Chang J, Collette NM, Ovcharenko D, Plajzer-Frick I, Rubin EM. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res* **15**: 928–935.

Losos JB, Creer DA, Glossip D, Goellner R, Hampton A, Roberts G, Haskell N, Taylor P, Ettling J. 2000. Evolutionary implications of phenotypic plasticity in the hindlimb of the lizard Anolis sagrei. *EVOLUTION* **54**: 301–305.

Ludwig MZ, Manu, Kittler R, White KP, Kreitman M. 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genet* **7**: e1002364.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Maeda YT, Sano M. 2006. Regulatory dynamics of synthetic gene networks with positive feedback. *J Mol Biol* **359**: 1107–1124.

Masel J, Siegal ML. 2009. Robustness: mechanisms and consequences. *Trends Genet* **25**: 395–403.

Masel J, Trotter MV. 2010. Robustness and evolvability. *Trends Genet* **26**: 406–414.

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. 2012. Widespread Site-Dependent Buffering of Human Regulatory Polymorphism ed. G.S. Barsh. *PLoS Genet* **8**: e1002599.

McPhail JD. 1992. Ecology and evolution of sympatric sticklebacks (Gasterosteus): evidence for a species-pair in Paxton Lake, Texada Island, British Columbia. *Can J Zool* **70**: 361–369.

Merika M, Thanos D. 2001. Enhanceosomes. *Curr Opin Genet Dev* **11**: 205–208.

Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, De Laat W, Spitz F, Duboule D. 2011. A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**: 1132–1145.

Nechaev S, Adelman K. 2011. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim Biophys Acta* **1809**: 34–45.

Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. 2011. The dynamic architecture of Hox gene clusters. *Science* **334**: 222–225.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. *Nature* **388**: 167–171.

O'Donnell AF, Tiong S, Nash D, Clark DV. 2000. The Drosophila melanogaster ade5 gene encodes a bifunctional enzyme for two steps in the de novo purine synthesis pathway. *Genetics* **154**: 1239–1253.

Oda H, Tsukita S, Takeichi M. 1998. Dynamic Behavior of the Cadherin-Based Cell–Cell Adhesion System during Drosophila Gastrulation. *Dev Biol* **203**: 435–450.

Ohler U. 2006. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res* **34**: 5943–5950.

Panne D, Maniatis T, Harrison SC. 2007. An atomic model of the interferon-beta enhanceosome. *Cell* **129**: 1111–1123.

Pappu KS, Ostrin EJ, Middlebrooks BW, Sili BT, Chen R, Atkins MR, Gibbs R, Mardon G. 2005. Dual regulation and redundant function of two eye-specific enhancers of the Drosophila retinal determination gene dachshund. *Development* **132**: 2895–2905.

Perry MW, Boettiger AN, Bothma JP, Levine M. 2010. Shadow enhancers foster robustness of Drosophila gastrulation. *Curr Biol* **20**: 1562–1567.

Perry MW, Boettiger AN, Levine M. 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci USA* **108**: 13570–13575.

Perry MW, Bothma JP, Luu RD, Levine M. 2012. Precision of Hunchback Expression in the Drosophila Embryo. *Current Biology* **22**: 2247–2252.

Perry MW, Cande JD, Boettiger AN, Levine M. 2009. Evolution of insect dorsoventral patterning mechanisms. *Cold Spring Harb Symp Quant Biol* **74**: 275–279.

Pigliucci M, Murren CJ, Schlichting CD. 2006. Phenotypic plasticity and evolution by genetic assimilation. *J Exp Biol* **209**: 2362–2367.

Polaczyk PJ, Gasperini R, Gibson G. 1998. Naturally occurring genetic variation affects Drosophila photoreceptor determination. *Dev Genes Evol* **207**: 462–470.

Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet* **26**: 425–430.

Raff RA, Sly BJ. 2000. Modularity and dissociation in the evolution of gene expression territories in development. *Evol Dev* **2**: 102–113.

Raj A, Rifkin SA, Andersen E, van Oudenaarden A. 2010. Variability in gene expression underlies incomplete penetrance. *Nature* **463**: 913–918.

Rau A, Buttgereit D, Holz A, Fetter R, Doberstein SK, Paululat A, Staudt N, Skeath J, Michelson AM, Renkawitz-Pohl R. 2001. rolling pebbles (rols) is required in Drosophila muscle precursors for recruitment of myoblasts for fusion. *Development* **128**: 5061–5073.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* **22**: 860–869.

Reim I, Frasch M. 2010. Genetic and genomic dissection of cardiogenesis in the Drosophila model. *Pediatr Cardiol* **31**: 325–334.

Reim I, Mohler JP, Frasch M. 2005. Tbx20-related genes, mid and H15, are required for tinman expression, proper patterning, and normal differentiation of cardioblasts in Drosophila. *Mech Dev* **122**: 1056–1069.

RENDEL JM. 1959. Canalization of the Scute Phenotype of Drosophila. *EVOLUTION* **13**: 425–439.

RENDEL JM, SHELDON BL. 1960. SELECTION FOR CANALIZATION OF THE SCUTE PHENOTYPE IN DROSOPHILA MELANOGASTER. *Aust J Biol* **13**: 36–47.

RENDEL JM, SHELDON BL, Finlay DE. 1965. Canalisation of development of scutellar bristles in Drosophila by control of the scute locus. *Genetics* **52**: 1137–1151.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–1419.

Rosenfeld N, Elowitz MB, Alon U. 2002. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* **323**: 785–793.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. *Nature* **396**: 336–342.

Saga Y, Yagi T, Ikawa Y, Sakakura T, Aizawa S. 1992. Mice develop normally without tenascin. *Genes Dev* **6**: 1821–1831.

Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EE. 2007. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* **21**: 436–449.

Sandmann T, Jensen LJ, Jakobsen JS, Karzynski MM, Eichenlaub MP, Bork P, Furlong EE. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**: 797–807.

Scharloo W. 1991. CANALIZATION: GENETIC AND DEVELOPMENTAL ASPECTS. *Annu Rev Ecol Syst* **22**: 65–93.

Schlosser G, Thieffry D. 2000. Modularity in development and evolution. Vol. 22 of, pp. 1043–1045.

Schluter D, McPhail JD. 1992. Ecological character displacement and speciation in sticklebacks. *Am Nat* **140**: 85–108.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci* **255**: 279–284.

Senger K, Armstrong GW, Rowell WJ, Kwan JM, Markstein M, Levine M. 2004. Immunity regulatory DNAs share common organizational features in Drosophila. *Mol Cell* **13**: 19–32.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458–472.

Siepel A. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Smale ST, Baltimore D. 1989. The “initiator” as a transcription control element. *Cell* **57**: 103–113.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449–479.

Small S, Blair A, Levine M. 1992. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* **11**: 4047–4057.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group* **13**: 613–626.

SSHASHIKANT C, RUDDLE FH. 1996. Combinations of closely situated cis-acting elements determine tissue-specific patterns and anterior extent of early Hoxc8 expression. *Proc Natl Acad Sci USA* **93**: 12364–12369.

Stanojević D, Small S, Levine M. 1991a. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **254**: 1385–1387.

Stanojević D, Small S, Levine M. 1991b. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **254**: 1385–1387.

Swanson CI, Evans NC, Barolo S. 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* **18**: 359–370.

Thanos D, Maniatis T. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**: 1091–1100.

Vavouri T, Semple JI, Lehner B. 2008. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet* **24**: 485–488.

Via S, Gomulkiewicz R, De Jong G, Scheiner SM, Schlichting CD, Van Tienderen PH. 1995. Adaptive phenotypic plasticity: consensus and controversy. *Trends Ecol Evol (Amst)* **10**: 212–217.

VISSER JAGMD, HERMISSON J, Wagner GP, MEYERS LA, BAGHERI-CHAICHIAN H, BLANCHARD JL, CHAO L, CHEVERUD JM, ELENA SF, FONTANA W, et al. 2003. Perspective: Evolution and detection of genetic robustness. *EVOLUTION* **57**: 1959–1972.

Vokes SA, Ji H, Wong WH, McMahon AP. 2008. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev* **22**: 2651–2663.

Waddington C. 1942. Canalization of development and the inheritance of acquired characters. *Nature* **150**: 563–565.

Waddington CH. 1953. Genetic assimilation of an acquired character. *EVOLUTION* 118–126.

WAGNER A. 2005. Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays* **27**: 176–188.

WAGNER A. 2008. Robustness and evolvability: a paradox resolved. *Proc Biol Sci* **275**: 91–100.

WAGNER A. 2012. The role of robustness in phenotypic adaptation and innovation. *Proc Biol Sci* **279**: 1249–1258.

Wagner CE, Harmon LJ, Seehausen O. 2012. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487**: 366–369.

Wagner G, Booth G, Bagheri-Chaichian H. 1997. A population genetic theory of canalization. *EVOLUTION* **51**: 329–347.

Wagner GP, Altenberg L. 1996. Perspective: Complex Adaptations and the Evolution of Evolvability. *EVOLUTION* **50**: 967–976.

West-Eberhard MJ. 2003. *Developmental Plasticity and Evolution*. Oxford University Press, USA.

Williams GC. 1966. *Adaptation and Natural Selection*. Princeton University Press.

Xiong N, Kang C, Raulet DH. 2002. Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity* **16**: 453–463.

Xu X, Yin Z, Hudson JB, Ferguson EL, Frasch M. 1998. Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the Drosophila mesoderm. *Genes Dev* **12**: 2354–2370.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865–2871.

Yi TM, Huang Y, Simon MI, Doyle J. 2000. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA* **97**: 4649–4653.

Yin Z, Xu XL, Frasch M. 1997. Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development* **124**: 4971–4982.

Zaffran S, Reim I, Qian L, Lo PC, Bodmer R, Frasch M. 2006. Cardioblast-intrinsic Tinman activity controls proper diversification and differentiation of myocardial cells in Drosophila. *Development* **133**: 4073–4083.

Zhang J. 2012. Genetic redundancies and their evolutionary maintenance. *Adv Exp Med Biol* **751**: 279–300.

Zichner T, Garfield DA, Rausch T, Stutz AM, Cannavo E, Braun M, Furlong EEM, Korbel JO. 2012. Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Res* 1–38.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**: 65–70.

Chapter 2

List of publications or manuscripts

Partial redundant enhancer elements are a major component of developmental robustness

Enrico Cannavò[1], Thomas Zichner[1], Pierre Khoueiry[1], E. Hilary Gustafson[1], Lucia Ciglar[1], Jan O. Korbel[1] and Eileen E. M. Furlong[1]*

[1]European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

*To whom correspondence should be addressed

E-mail furlong@embl.de

Telephone +49 6221 3878416

Fax +49 6221 3878166

Running title: Enhancer redundancy and transcriptional robustness

Keywords: Enhancer, redundancy, robustness, transcriptional networks, development

Chapter 3

Introduction

The synthesis of RNA is regulated at different levels from several elements in *cis* or in *trans*. The DNA sequence must be accessible to allow the transcriptional machinery to be recruited and assembled upstream of the transcription start site (TSS). Transcription is initiated by the binding of transcription factors at enhancers, which is an essential first step to activate saptio-temporal transcription. There are also many post-transcriptional regulatory steps, such as 5'-capping, splicing and 3'-polyadenylation and export from the nucleus that are important for correct gene expression. In the previous section of this thesis, I described events that concern the recruitment of RNA polymerase II (Pol II) to the promoter and the role of transcription factors in the activation of transcription. I have described how developmental reactions reach robustness through several mechanisms including the presence of multiple partially redundant enhancers: they confer genetic and environmental robustness to a canalized trait. In this section I will focus on the characterization of RNA termini, via polyadenylation in a wild type population of *D. melanogaster*.

3' polyadenylation

Importance of the 3' polyadenylation

In eukaryotes, messenger RNAs (mRNAs) are transcribed by RNA polymerase II. The mRNA precursors undergo extensive co-transcriptional processing before they can be transported to the cytoplasm: 5' capping where the guanosine at the most 5' end is methylated; splicing where intronic sequences are removed from the premature mRNA and polyadenylation, a co-transcriptional endonucleolytic cleavage that is followed by addition of a series of adenosines. Polyadenylation depends on multiple *cis* elements that are recognized by several multisubunit protein complexes.

mRNA polyadenylation is very important in eukaryotes and its alteration is the cause of several human diseases (Elkon et al. 2013). 3' end polyadenylation promotes mRNA transport from the nucleus to the cytoplasm (Huang and Carmichael 1996; Vinciguerra and Stutz 2004; Ji et al. 2011) and mRNA stability (Ford et al. 1997; Zhang et al. 2010; Bernstein et al. 1989). Indeed *in vitro* studies have shown that the addition (Ford et al. 1997) and length (Zhang et al. 2010) of the poly(A) tail, as well as the binding of the poly(A) binding protein (PABP) (Bernstein et al. 1989) have been associated with increased mRNA stability. The PABP indirectly interacts with the 5'cap through the translation initiation complex eIF4G which, in turn, forms a complex with the cap binding protein eIF4E (Wells et al. 1998). *In vitro* experiments showed that the interaction between the 5' and the 3' mRNA ends causes circularization and increases the stability of the mRNA, preventing access of deadenylating and decapping enzymes to their targets (Wells et al. 1998; Wilusz et al. 2001). Moreover, biochemical studies demonstrated that the carboxy-terminal domain of Pol II supports 3' processing by interacting with some of the protein complexes of the polyadenylation machinery in order to reach maximum levels of polyadenylation *in vivo* (Fong and Bentley 2001; McCracken et al. 1997). The polyadenylation machinery interacts with the activator transcription factors (TFs) in human and yeast: this led to the proposition that TFs function as anti-terminators, inhibiting the termination activity of the polyadenylation machinery during transcription elongation (Calvo and Manley 2001). Alterations of these interactions cause improper polyadenylation (Calvo and Manley 2001; Fong and Bentley 2001). Additionally, a number of studies indicate that the 3' polyadenylation and splicing process are tightly coupled so that, for example, splicing of the last intron is enabled (Martinson 2011).

A recent work showed how antisense polyadenylated transcripts at the TSS of active genes enforce promoter directionality (Ntini et al. 2013). In particular it was observed that unfavorable chromatin environment upstream of the TSS could favor elongation in one direction only. Ntini *et. al.* showed a unsymmetrical distribution of poly(A) sites that could contribute to confer directionality to transcription (Ntini et al. 2013).

Polyadenylation: motifs and subcomplexes

The 3’ polyadenylation process is directed by sequence elements in 3’ untranslated regions (UTR) of the premature mRNA. They are found in every eukaryotic polyadenylated mRNA and their disruption of position and sequence causes reduction in processing efficiency. Although these elements do differ between mammals, yeast and plants, there is a common tripartite arrangement that includes a polyadenylation signal (PAS), a cleavage site and a G/U rich downstream element (DSE) (Wahle and Keller 1992).

The PAS is a proximal element generally localized about 10-50 bp upstream the poly(A) site (Fig. 12) (Elkon et al. 2013; Zhao et al. 1999). In humans, this AAUAAA hexamer and its variant AUUAAA are the most frequent poly(A) motifs, in 58.2% and 15% of 3’ fragments respectively, while the other 9 variants collectively represent about the 14% of the sites (Beaudoing et al. 2000). Quantitative studies showed that the different variants of PAS are associated with different poly(A) strength (Yoon et al. 2012). The second element of the core polyadenylation signal, the downstream element (DSE), is within approximately 40-100 nucleotides downstream of the poly(A) site. This is the least conserved motif and two main types have been described: a short run of uridine, the U-rich type, and the GU-rich type with a UGUGU consensus sequence (Fig. 12) (Mandel et al. 2008; Zhao et al. 1999; Elkon et al. 2013). The third element is the cleavage site positioned between the PAS and the DSE. The cleavage site is not really conserved but one study in humans and yeast showed that it is located at the 3’ side of an adenosine within a region of 13 nucleotides in 70% of cases (Mandel et al. 2008; Chen et al. 1995). Auxiliary upstream and downstream motifs such as the UGUA sequence or the upstream sequence element (USE) have also been identified (Fig. 12); in some cases they enhance 3’ processing (Proudfoot 2011; Ozsolak et al. 2010).

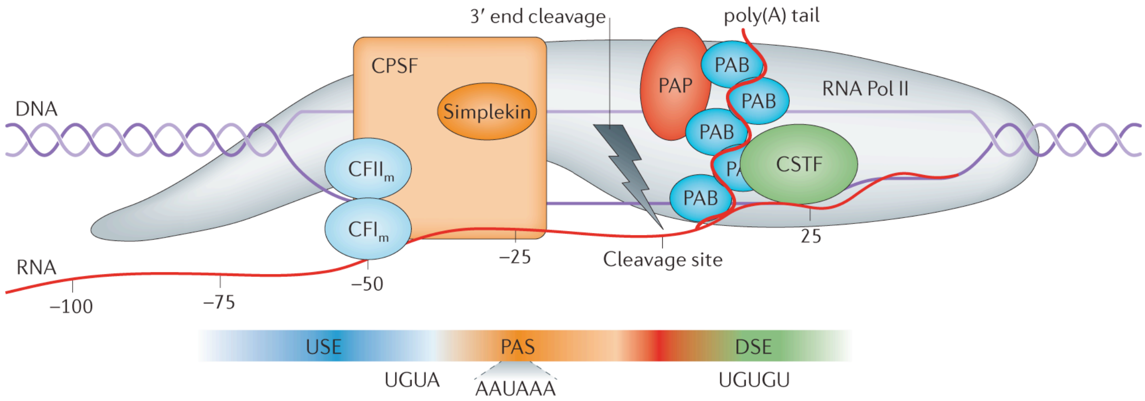


Fig. 12 Core complexes of the polyadenylation machinery. Cleavage and polyadenylation requires several *cis*-acting RNA elements and several dozen core and auxiliary polypeptides: Cleavage and polyadenylation specificity factor (CPSF), cleavage stimulating factor (CSTF), cleavage factor Im (CFIm) and CFIIm that bind respectively the polyadenylation signal (PAS), the downstream element (DSE) and the upstream element (USE). All polyadenylation factors are required for the cleavage and the polyadenylation through the recruitment of the poly(A) polymerase (PAP). The poly(A) tail is then bound by the poly(A) binding protein (PAB). Image from (Elkon et al. 2013).

The polyadenylation machinery contains more than 14 proteins forming several sub-complexes. In particular, the cleavage and polyadenylation specificity factor (CPSF) binds with high affinity to the canonical PAS AAUAAA and with less affinity to all the other variants (Fig. 12) (Zhao et al. 1999). One of its subunits (CPSF73) is responsible for the cleavage of the mRNA at the cleavage site (Mandel et al. 2006). CPSF interacts with the pre-initiation complex (PIC), in particular with TFIID and with the C-terminal domain (CTD) of Pol II (Fong and Bentley 2001; McCracken et al. 1997; Dantonel et al. 1997). The efficient binding of CPSF also depends on interaction with the cleavage stimulation factor (CstF). CstF interacting with the DSE is important for the proper cleavage (Fig. 12) (Mandel et al. 2008). Like CPSF, CstF is recruited to the transcription PIC by interacting with transcription factors and the CTD of Pol II (Calvo and Manley 2001; McCracken et al. 1997). Another important component in the 3’ processing machinery is the mammalian cleavage factor I (CFIm). CFIm binds to the

USE and is important in the recognition of the poly(A) signal when the perfect PAS is absent, as well as in enhancing the binding of CPSF and CstF to their motifs (Fig. 12) (Venkataraman et al. 2005). Together with the mammalian cleavage factor II (CFIIm), the CFIm is important for the cleavage of the pre-mRNA (Zhao et al. 1999). Concerning the PAP, it was demonstrated that it is required for polyadenylation *in vivo,* while *in vitro* it also mediates the cleavage (Zhao et al. 1999). The interactions of the PAP with the other factor of the polyadenylation machinery regulates the length of the poly(A) tail (Mandel et al. 2008). The polymerization of the adenosine tail to a specific length depends on the species: for instance it is 150–250 nucleotide long in mammals (Brown and Sachs 1998). Finally the poly(A) binding protein (PAP) binds the nascent poly(A) tail as it becomes available until the proper length is reached (Fig. 12) (Meyer et al. 2002). The CTD of Pol II also plays an active role in the recruitment of polyadenylation factors, such as CPSF, CstF and one of the subunits of CFIIm, to the Pol II elongation complex (Zhao et al. 1999; McCracken et al. 1997; Meinhart and Cramer 2004). Indeed truncation of the Pol II CTD showed aberrant and inefficient polyadenylation *in vivo* (McCracken et al. 1997).

Thus, polyadenylation of most coding genes requires the recruitment and synergistic action of core and associated proteins. The initiating step is the interaction of CPSF and CstF with the transcription PIC. Upon transcriptional initiation, the CTD of Pol II is phosphorylated, the PIC is released and polyadenylation components associate with the CTD. When Pol II approaches the end of the transcript, the polyadenylation factors recognize the respective motifs on the precursor mRNA. The individual interactions of CPSF and CstF with their motifs are weak but are stabilized by their cooperative binding in a process assisted by CFIm. These bindings events define the region where the cleavage site lies. CFIm, CFIIm and the PAP are then recruited, the cleavage is enabled and it is followed by the polymerization of the poly(A) tail.

The only protein-encoding genes that lack a poly(A) tail are the replication-dependent histone genes where the mature 3' end depends on significantly different *cis* and *trans* components from the polyadenylation process (reviewed in Proudfoot 2004).

The process of polyadenylation and splicing are closely coordinated. Such connection has been shown in the recognition of the terminal exon, where the downstream 3' splice factors enhance polyadenylation (Martinson 2011). The recognition of the 3' splice site (3'SS) of the last intron of a gene enhances the efficiency of its polyadenylation. In particular the polypirimidine tract is bound by the U2AF factor which, in turn, interacts with CFIm complex and with the poly(A) polymerase (Martinson 2011). Moreover the U2 snRNPs that associate with the 3'SS, interacts with the polyadenylation factor CPSF (Martinson 2011). Another splicing factor that interacts with CPSF is SRm160, a splicing co-activator that helps to create a bridge during the intronic definition (McCracken et al. 2002). All these interactions serve to stabilize the assembly of the polyadenylation machinery at the end of the transcript.

Alternative polyadenylation

Recently, it has become evident that mRNAs have multiple alternative 3' ends that are formed by cleavage and polyadenylation at different sites. This phenomenon, called alternative polyadenylation (APA), has been extensively observed in several species. For example, in mammals, *S. cerevisiae* and *Arabidopsis* ~70% of expressed genes have APA (Shi 2012; Derti et al. 2012; Ozsolak et al. 2010) while in *D. melanogaster* only 55% were reported to have APA (Smibert et al. 2012). In general, two categories of APA have been described: either they are found in internal introns/exons, producing a different protein isoform or they are located in the same 3' UTR resulting in several 3'UTR with different length that do not effect the coding of the protein, but which may include different regulatory motifs (Tian et al. 2007; 2005; Beaudoing et al. 2000). Many terms have been used to define these two classes of APA and in some cases they have been subdivided in even more classes (Di Giammartino et al. 2011; Shi 2012; Jan et al. 2011): here I distinguish between only these two classes and I refer to the first class as exonic/intronic APA (EI-APA) and to the second class as UTR-APA.

The global analyses of 3' processing in multiple species allowed the evolutionary conservation of APA to be examined. It has been shown that nucleotide composition, core *cis* elements and their positions are well conserved among metazoa. In general, the canonical PAS AAUAA is the most common among all described variants (Jan et al. 2011; Ozsolak et al. 2010; Shi 2012). The palindromic sequences of A-rich and U-rich motifs has been observed to behave as double duties in cases where two convergent genes are in opposite strands so that the same poly(A) motif is used by two neighboring genes (Jan et al. 2011; Ozsolak et al. 2010). Overall the distal PAS tend to have canonical strong consensus PAS and strong DSE in contrast to weaker non-canonical proximal UTR-APA and EI-APA sites (Smibert et al. 2012; Beaudoing et al. 2000; Martin et al. 2012; Shepard et al. 2011; Tian et al. 2007). Moreover the length of the 3' UTR has been linked to the proliferation, cell type and cell differentiation state (Ji et al. 2009; Ji and Tian 2009; Derti et al. 2012; Smibert et al. 2012), suggesting that distal PAS could be used as default polyadenylation sites while the proximal PAS could play a regulatory role (Shi 2012).

Different mechanisms regulate the choice of APA sites. One way to regulate the APA is by controlling the expression level of general polyadenylation factors (Takagaki et al. 1996; Chuvpilo et al. 1999). A well-characterized example of this model is represented by B cell differentiation. CstF-64 is a subunit of the CstF complex that directly contacts the DSE region of the pre-mRNA (Mandel et al. 2008). Induction of B cells with lipopolysaccharide (LPS) increases the protein level of CstF-64 (Takagaki et al. 1996) that, in turn, causes the switch of the immunoglobin M (IgM) heavy chain from a membrane-bound isoform to a secreted isoform. This switch is determined by a different usage of poly(A) sites, from a distal to a proximal site because of different affinities of two DSEs: in case of low level of CstF-64 expression it binds a strong high affinity downstream DSE while at higher concentrations it binds a weaker and more upstream DSE (Takagaki et al. 1996). In general, the expression level of polyadenylation factors is higher in proliferating and differentiating cells (Ji and Tian 2009; Elkon et al. 2012): this regulation can be transcriptionally controlled by the transcription factor E2F (Elkon et al. 2012) or by signaling pathways such as the one of p38 MAP kinase activated by stress conditions (Danckwardt et al. 2011).

APA can be modulated by positive or negative regulatory factors, some of which compete with the polyadenylation factors for binding to their specific motif. Sex lethal is a RNA binding protein that is important in regulating the splicing and translation of gender specific transcripts in *D. melanogaster*. This protein competes with CstF-64 for binding to the DSE motif in the female germline causing a switch in the usage of UTR-APA from proximal to distal (Gawande et al. 2006). Many other cases of splicing factors that control APA have been described (Castelo-Branco et al. 2004; Moreira et al. 1998; Danckwardt et al. 2007; Millevoi et al. 2009). The polypyrimidine tract binding factor (PBT) is a major heterogeneous nuclear ribonucleoprotein (hRNP) that regulates splicing in metazoans. PBT regulates 3' processing by competing with the binding of CstF to the DSE motif, thus inhibiting 3' mRNA cleavage (Castelo-Branco et al. 2004). In other cases, the PBT can enhance mRNA polyadenylation by recruiting hnRNPs, i.e. Nova2, to auxiliary polyadenylation elements (Danckwardt et al. 2007; Millevoi et al. 2009). Tissue specific splicing factors can also modulate the 3' process: Nova2 is a mouse RNA binding protein that can positively or negatively regulate the polyadenylation process in the mouse brain depending on the position of its binding site (Licatalosi et al. 2008). These data suggest a clear interaction between the polyadenylation machinery and splicing factors that compete for the same motifs inhibiting or enhancing the recruitment of polyadenylation machinery at the 3' end of transcripts.

Moreover, splicing factors can modulate the usage of polyadenylation sites through the binding to sites outside the polyadenylation motifs. A well-characterized example of this modulation is represented by U1 snRNP. U1 snRNP plays an essential role in defining the 5' splice site by RNA:RNA base pairing via U1 snRNA's 5' nine-nucleotide sequence (Wahl et al. 2009). This binding globally suppresses premature cleavage and polyadenylation within intronic regions in a dose dependent manner (Kaida et al. 2010; Berg et al. 2012). Physiologically, neuronal activation results in a 40%–50% increase in nascent transcripts, while the concentration of U1 snRNP shows little change in expression level. It was suggested that this increase in transcription could create a significant U1 shortage: this could be the cause of the observed short 3'UTRs in neuronal cells, as many neuronal genes used the proximal PAS motif (Berg et al. 2012).

As previously mentioned, the polyadenylation machinery associates with the transcriptional PIC. Several studies suggest that transcription factors can modulate the recruitment of polyadenylation factors as well as alternative polyadenylation. CPSF and CstF are independently recruited by general TFs to genes promoters (McCracken et al. 1997; Dantonel et al. 1997). During transcriptional elongation, they interact with the CTD of the Pol II elongation complex which, in turn, is necessary for efficient 3' processing (McCracken et al. 1997; Dantonel et al. 1997). ChIP assays have shown that polyadenylation factors colocalize both at the 5' and 3' of genes and that this is associated with Pol II pausing (Glover-Cutter et al. 2008). In particular, pausing of Pol II is more likely at the polyA site of highly expressed genes than lowly expressed genes (Luo et al. 2011). It was also observed that through the MED13 subunit, the mediator influences the APA profile regulating the hnRNP L occupancy or interacting directly with the polyadenylation factors (Huang et al. 2012). All these data show that general transcription factors promote the efficient recruitment of polyadenylation factors and the recognition of polyadenylation motif.

The human Paf1C is a factor that has a central role in orchestrating cotranscriptional histone modifications required for histone H2B monoubiquitination (Rozenblatt-Rosen et al. 2009). It is also an elongation factor that acts synergistically to enhance elongation together with another elongation factor, TFIIS (Nagaike and Manley 2011). *In vitro* experiments showed that one of the Paf1C subunits interacts with Pol II and recruits polyadenylation factors enhancing the usage of proximal polyadenylation sites (Rozenblatt-Rosen et al. 2009). Indeed a recent work emphasizes the potential connection between the transcription elongation rate and APA: Pol II mutants with slower elongation kinetics favor the recognition of the proximal polyadenylation sites in contrast with cells expressing wild type Pol II where the usage of distal polyadenylation sites increases. How elongation kinetics influences the APA and the splicing is poorly understood. Evidence coming from recent studies suggest that chromatin and chromatin-associated factors could play an important role. In particular strong nucleosome depletion around polyadenylation sites and nucleosome enrichment just downstream of those sites were observed (Spies et al. 2009): highly used APA sites had higher downstream nucleosome density than less frequently used alternative sites. These differences suggest that nucleosomes positioning might influence PAS usage through effects on the kinetics of polymerase elongation in the vicinity of the PAS (Spies et al. 2009). Epigenetic modifications can also influence utilization of alternative polyadenylation sites: the proximal PAS in genes expressed at high levels tend to have low nucleosome and higher H3K4me3 and H3K36me3 levels were observed (Ji et al. 2011). Moreover, imprinted regions and methylation of CpG islands can influence the usage of proximal or distal APA sites depending on the transcription rate of Pol II (Wood et al. 2008).

Together with alternative splicing, APA can increase the repertoire of transcripts produced by a gene, expanding the proteomic diversity. The case of the IgM gene discussed previously is a clear example. Here, different polyadenylation sites can influence and change the expressed of protein isoforms: for example, in another case, polyadenylation in a constitutively spliced internal exon of the *EPRS* gene changes the coding sequence and consequently the isoform of the expressed protein. This truncated protein attenuates the function of the full-length protein acting as a dominant-negative inhibitor that prevents complete translational silencing of target transcripts (Yao et al. 2012). Polyadenylation factors also regulate their own expression stimulating the usage of alternative polyadenylation sites that can cause the expression of truncated and unstable proteins (Dai et al. 2012; Mansfield and Keene 2012).

The most obvious consequence of variation in the length of a transcripts UTR is the presence or absence of *cis* regulatory motifs. A systematic examination of the UTR-APA found that 52% of microRNA target sites are located downstream of the proximal poly(A) site (Legendre et al. 2006). Indeed proliferating murine T cells have shorter 3'UTRs and forced expression of full-length 3′UTRs conferred reduced protein expression that in some cases could be reversed by deletion of the predicted microRNA target sites (Sandberg et al. 2008). Shortening the 3' UTR has been observed in cancer cells, which leads to an increased stability of mRNAs and consequently to higher proteins levels, in part because of a loss in microRNA-mediated repression (Mayr and Bartel 2009).

APA can influence protein expression by effecting the localization of the mRNA. BDNF is a neutrophic factor that plays a plethora of functions in the brain. Its transcript has two isoforms, with a short or long 3'UTR that are differently regulated in translation (Lau et al. 2010). The short 3'UTR BDNF mRNA is associated with the polyribosome fraction and mediates active translation to maintain basal levels of BDNF protein production, while the long isoform is sequestered into translationally dormant ribonucleoprotein particles in dendrites. Upon activation, the long 3'UTR BDNF mRNA is released and translated into both the somatal and dendritic compartments so that the protein level increases but the mRNA level remains stable (Lau et al. 2010).

All these studies showed that APA has a crucial role in the regulation of gene expression in several conditions and mostly in a tissue specific way. Here, we extensively optimized a quantitative and high sensitive 3’TagSeq protocol that allowed us to to study the regulation of APA and their differential temporal usage in a large population of wild-type *D. melanogaster* developing embryos. Due to the quantitative properties of the 3’TagSeq protocol, we will then use these data to investigate the role of genetic perturbation on the regulation of developmental gene expression.

Results

Examining variation in mRNA poly adenylation during embryonic development

To study variation in gene expression in *D. melanogaster,* we collected embryos from several isogenic fully sequenced lines from the *Drosophila melanogaster Genetic Reference Panel* (DGRP) (Mackay et al. 2012). These lines have been generated from a heterogeneous wild type population from Raleigh, North Carolina, USA, and have been self-crossed for at least 20 generations to obtain a set of highly homozygous inbred lines (Mackay et al. 2012). Depending on the quality of their genome sequencing we selected 80 of those lines for this study.

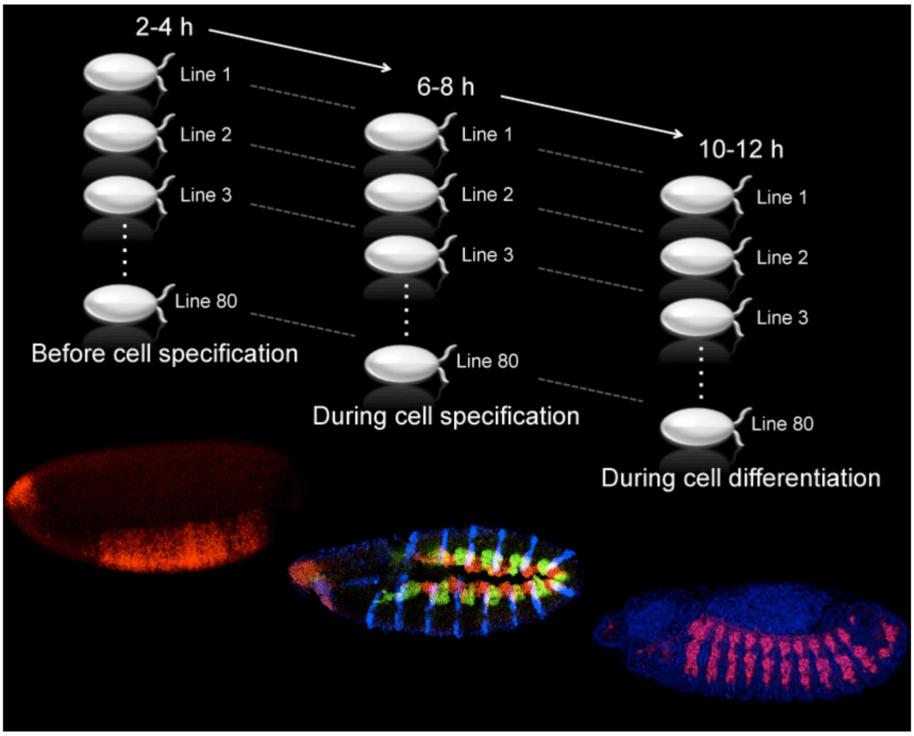


Fig. 13 Schematic representation of the embryo collections. Embryos from 80 DGRP lines where collected at three different time point. All embryos at the indictaed stages are oriented anterior - left, dorsal – up.

I analyzed their expression profiles by collecting embryos at three different embryonic windows: stage 5-8 (when cells are still multipotent), stage 11-12 (during cell specification), stage 13-15 (during cell differentiation). The samples were collected in at least three biological replicates per line and per time point for a total of more than 1,000 collected samples.

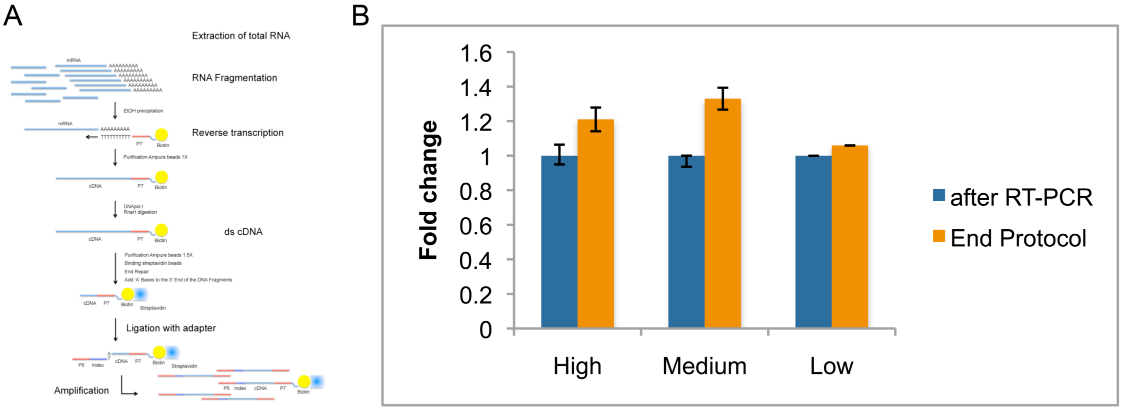


Fig. 14 Optimization of the 3'TagSeq protocol. A Scheme of the protocol. **B** qPCR on three gene expressed at different level (high, medium and low) across embryonic development. Comparisons of the expression level between the first step of the protocol and the level of genes at the end of the protocol confirmed the quantitative level of differentially expressed gene is maintained.

I focused our attention on the mechanisms and regulation of 3' mRNA processing. For this, I extensively modified and optimized previously published strand specific 3' TagSeq methods (Wilkening et al. 2013) to facilitated a quantitative analyze of Alternative PolyAdenylation (APA) in wild type *D. melanogaster* individuals during embryonic development (Fig. 14 A). First of all, I demonstrated that our protocol is highly specific since I could selectively detect signals coming from the end of mRNAs, including APA, as described before (Smibert et al. 2012). Through qRT-PCR experiments I confirmed that the protocol is highly quantitative and that it does not introduce biases that could alter the quantitative level of differentially expressed genes, even for low abundance transcripts (Fig. 14 B). Furthermore, to assess the performance of the 3'TagSeq protocol, I compared the results obtained with our 3'TagSeq method to standard RNA-Seq observing a significantly high correlation (Spearman' s rho= 0.84) (Fig. 15 A).

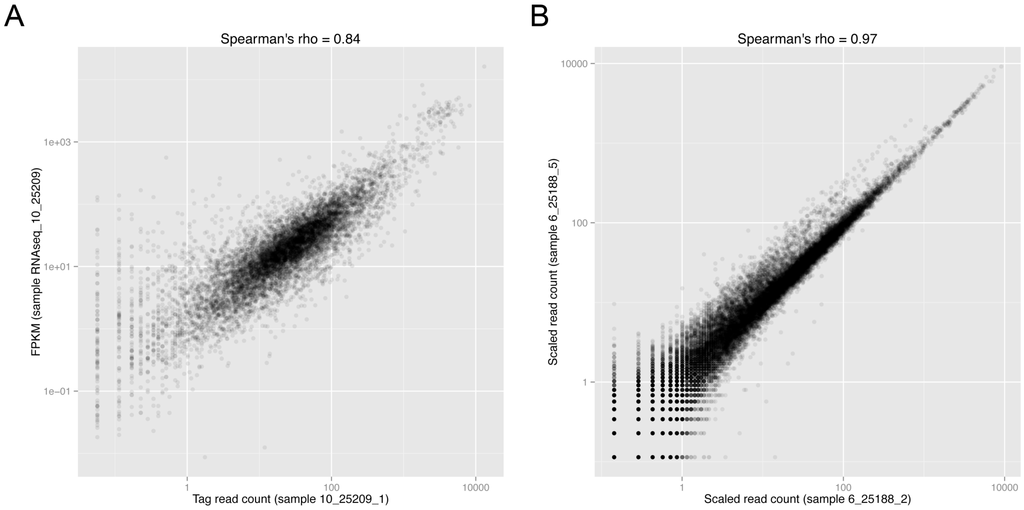


Fig. 15 The 3'TagSeq protocol is highly reproducible. A Comparison 3'TagSeq and RNA-Seq using the same RNA sample for one line (25209) from 13-15 stage embryos. **B** Biological replicates of the 11-12 stage embryos from the line 25188 show high reproducibility of the protocol.

The method is both highly reproducible, giving a typical rho = 0.97 (Fig. 15 B) correlation between biological replicates, and sensitive, being able to detect expression for genes expressed in only about 100 cells per embryo (i.e. less 0.5% of cell in the embryo), as exemplified by *tinman*.

In summary, my modified 3'TagSeq protocol has high specificity, sensitivity and reproducibility allowing me to perform the first quantitative assessment of variation in alternative polyadenylation during embryonic development.

Using the 3' TagSeq protocol, I assessed alternative poly adenylation of RNA transcripts for 80 lines at different stages of development, yielding a total of 246 datasets. To exclude any errors from inadvertently mixing or incorrectly labeling samples from a given genotype at any steps of the experiment, I evaluated the identity of each of the 80 lines at specific loci using diagnostic SNPs and more globally using the transcript data chromosome 2.

Often, the process of being inbred can cause differences in the time it takes for development to reach completion, such as developmental delays. If this occurs in one line (or ‘individual’), it would result in what appears to be many differentially expressed transcripts in that individual compared to the other lines. To reduce the possibility of having large expression changes due to temporal shifts in development, rather than having effects on a small number of transcripts due to specific SNPs, we correlated expressed genes in our dataset to that of RNA-Seq data from an already published time course of expression during embryonic development (Fig. 16) (Graveley et al. 2011).

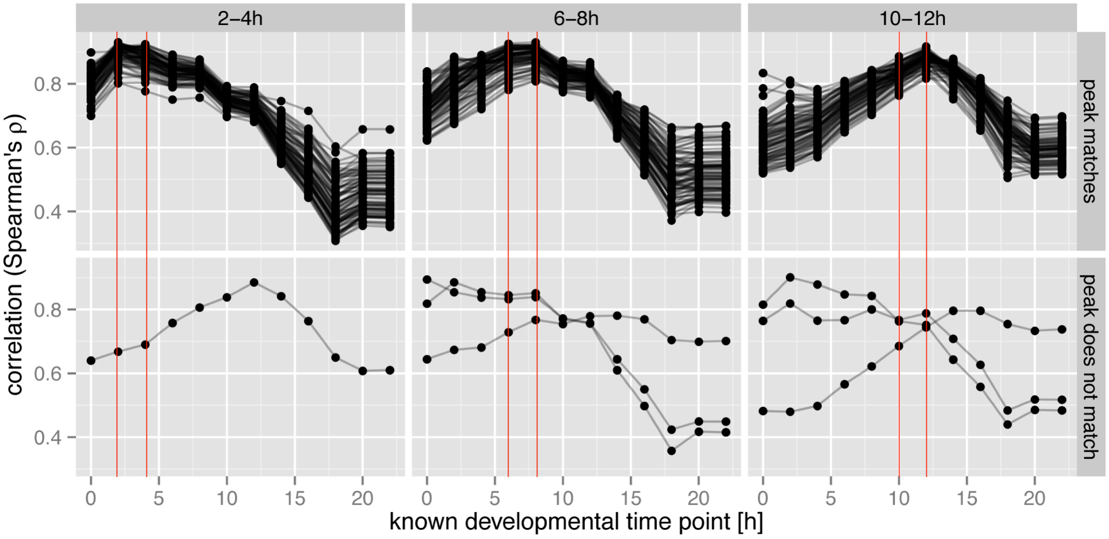
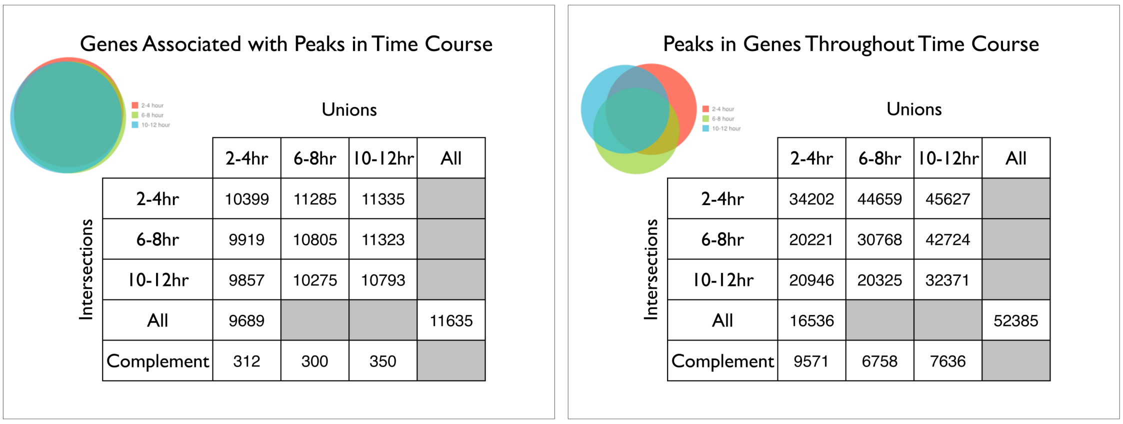


Fig. 16 Control on the embryonic stage. Comparison between an embryonic time course RNA-Seq and 3’TagSeq data to rule out the hypothesis of mis-staged collected embryos from the 80 different lines. We observe only a minority of samples at the non-appropriate stage (bottom pannel).

In general out of the 256 sequenced libraries, we detected 5% ‘errors’ due to mis-staging and the wrong genotype of the sample. In each of these cases, the problem has been fixed by reassigning the correct genotype and in the worst case, by discarding the data.

Defining the 3’end of transcripts

After having assessed the quality of the protocol and the sequenced data, we focused our attention on the 3’ poly adenylation process. In general we observed about 10,000 expressed genes per developmental window (Table. 1).



2-4 hour
6-8 hour
10-12 hour

Genes Associated with Peaks in Time Course

Unions / Intersections

	2-4hr	6-8hr	10-12hr	All
2-4hr	10399	11285	11335	
6-8hr	9919	10805	11323	
10-12hr	9857	10275	10793	
All	9689			11635
Complement	312	300	350	

Peaks in Genes Throughout Time Course

Unions / Intersections

	2-4hr	6-8hr	10-12hr	All
2-4hr	34202	44659	45627	
6-8hr	20221	30768	42724	
10-12hr	20946	20325	32371	
All	16536			52385
Complement	9571	6758	7636	

Table 1 Expression analysis of the three developmental windows. Left: genes that are expressed at specific stages. Most genes are expressed across the developmental windows analyzed. Few hundred are expressed in a stage specific way. **Right**: Peaks in genes show more dynamics. More than 6500 peaks are stage specific, indicating genes are dynamically expressed at different level through the development.

The bulk of the ~700 million high quality reads, pooled from all 80 sequenced lines at all three time points, localized at the 3’ end of annotated genes (Fig. 17). Although expected, this again confirms the quality and specificity of the 3’ Tag-Seq protocol. In addition to this, we however also detected a smaller minority of reads at the TSSs, in exons and introns of expressed genes (Fig. 17), in agreement with previous studies describing genome-wide APA (Smibert et al. 2012b; Beaudoing et al. 2000; Martin et al. 2012; Shepard et al. 2011; Tian et al. 2007). We identified more than 30,000 poly(A) peaks per time points (total of 52,385 peaks for all analyzed developmental stages). Importantly, 100% of all expressed genes have at least one poly(A) peak, with 35% of having a single poly(A) peak at their 3’ end. Interestingly,

~65% of all expressed genes have more than one poly(A), with some genes having up to 180 (Fig. 17).

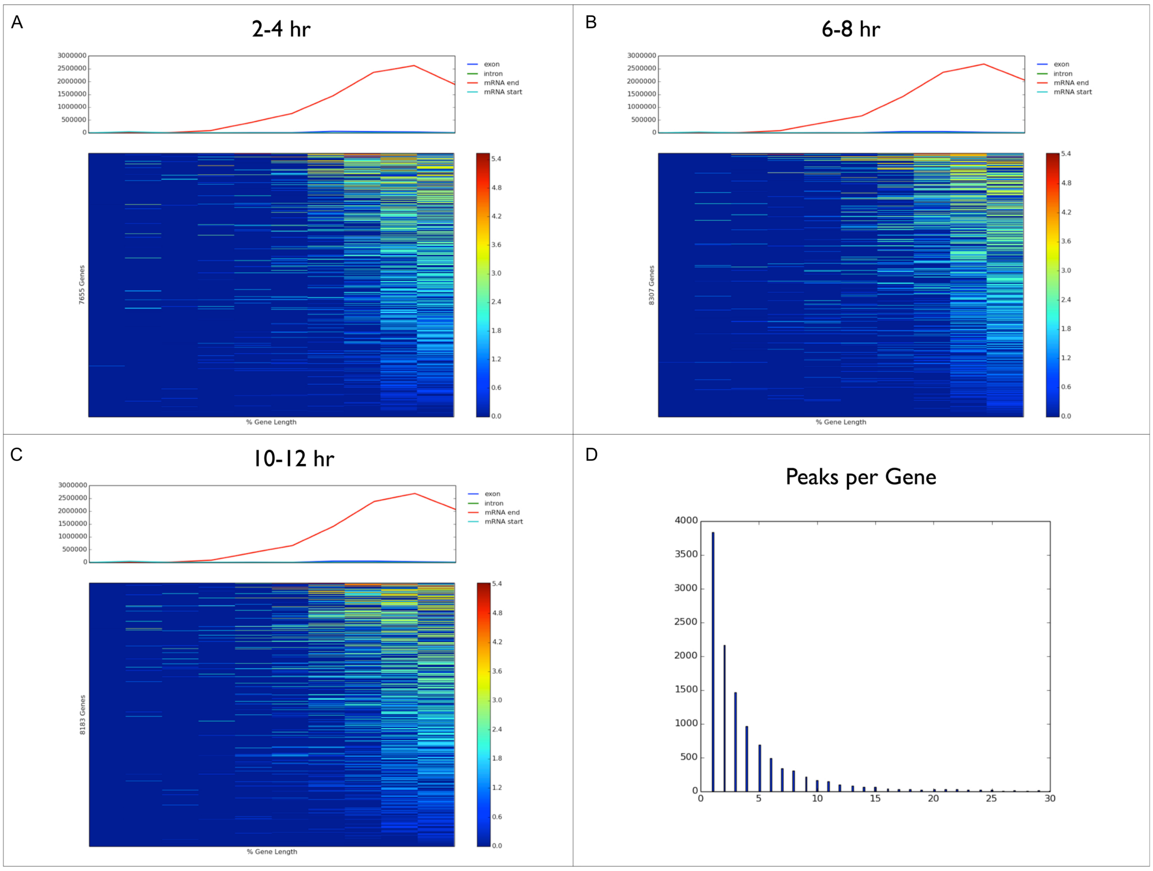


Fig. 17 Reads localize mostly at the 3'end of genes. A-C Most of 700 Million reads localize at the 3' end of the genes (red line) in comparison to TSS (light blue), exonic (dark blue) and intronic (green) regions at all three time points. **D** Peaks per gene show that APA is more pervasive than previously shown.

As expected, these peaks localize mostly at the 3' of the gene (about 60%), while fewer peaks are around TSS, in exonic and intronic regions (2%, 8%, 5%, respectively). Interestingly, we also detected a significant number of peaks (4.6%) in non-annotated gene regions across the analyzed developmental stages.

Taken together, we have defined a high-quality single-base resolution polyadenylation map for 80 'individuals' during three key stages of embryonic development. Our results indicate that APA is more pervasive than previously reported

(Smibert et al. 2012a), and is much more in line with the extent of APA described in other species (Shi 2012; Derti et al. 2012; Ozsolak et al. 2010).

Polyadenylation motifs

A number of sequence motifs are involved in the regulation of the polyadenylation process. Having precisely defined the poly(A) peaks of all transcripts with single nucleotide resolution, we used *de novo* motif discovery to identify upstream and downstream motifs linked with polyadenylation. This analysis provided two useful insights. First, it identified all previously described motifs associated with polyadenylation factors, unscoring the quality and resolution of the data: the most enriched motif is the canonical AAUAAA with its 10 variants, which is located 50 bps upstream of the end of the transcripts (Fig. 18) (Beaudoing et al. 2000). The DSE motif is also enriched at the expected position (within 25 nucleotide downstream the cleavage site) (Fig. 18). Second, it identified additional auxiliary motifs in different positions than previously described, such as poly-timidine or poly-adenine motifs. We observed that stretches of adenine sequences both just upstream and downstream (from -20 to +15 bp) of the TES and in between the PAS and DSE (Fig. 18). This suggests that this motif can be bound by the PAP and could play a role in the regulation of the cleavage process. Conversely, we found that the poly(T) motif is located downstream the TES (from +1 to +25 nucleotides) differently than previous studies where it was observed within 21 nucleotides upstream the TES (Fig. 18). Interestingly, the USE motif is located both at expected positions (from -40 to -75) and just upstream and downstream the TES (Fig. 18).

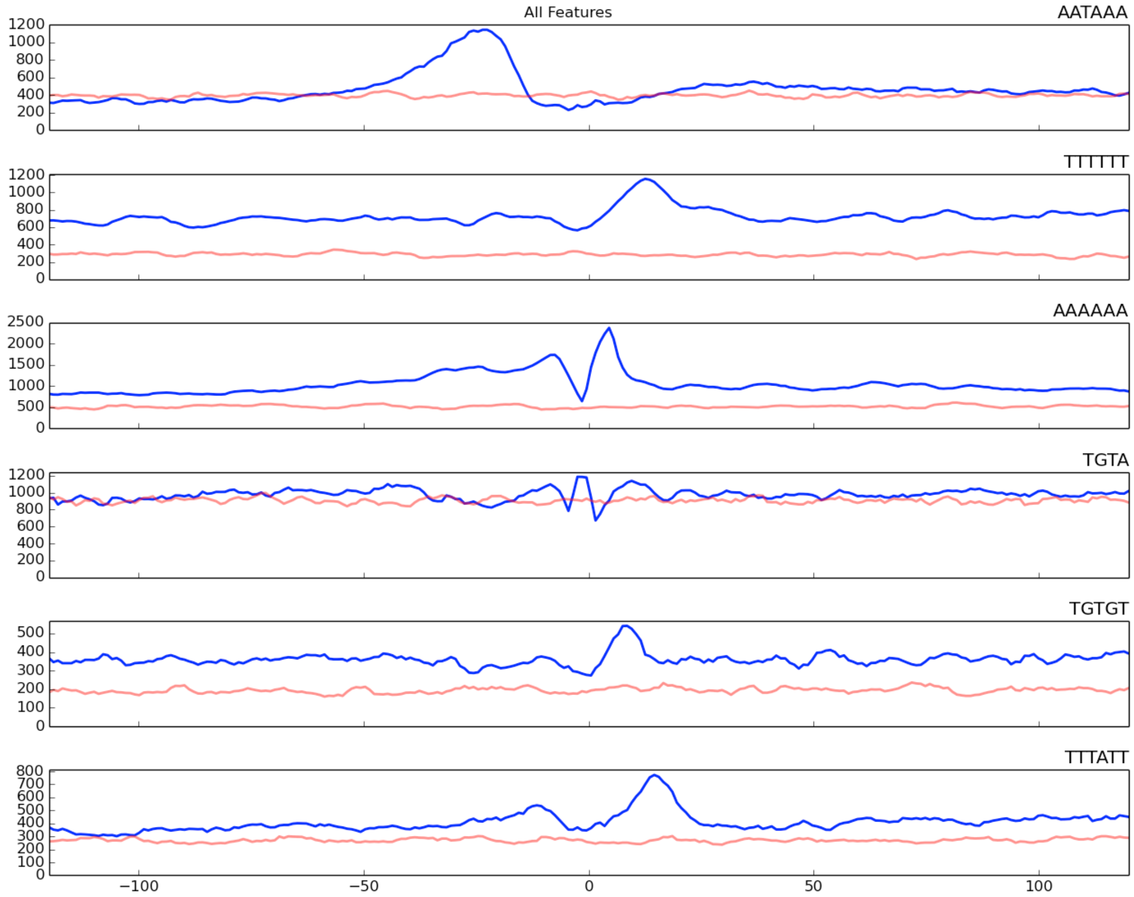


Fig. 18 ***de novo*** **polyadenylation motifs identification.** Identification of the canonical PAS, DSE, USE and other auxiliary sequences from the 3' TagSeq peaks from all 80 lines.

Since we observed poly(A) sites not only at the annotated 3' end of transcripts but also in exons, introns and TSS, we next analyzed if a subset of motifs is specifically enriched in peaks that are located in any of these spatially regions of a gene (e.g. specific to the TSS versus the 3' end etc) (Fig. 19). In all seven classes examined, the most enriched motif is the canonical upstream PAS (AAUAAA), the upstream poly(A) motif and the downstream poly(T) sequence. As expected, most of the upstream and downstream motifs are significantly enriched at the annotated mRNA end (Fig. 19). Importantly, we noticed that some motifs are differentially enriched in a specific class: for example, one of the variants of the PAS is enriched in exonic poly(A) peaks. We also identified a reverse complement motif to the canonical polyadenylation sequence, TTTATT, that is enriched 25 nucleotides upstream the TES of poly(A) peaks, within 2 kb downstream the annotated end of genes (Fig. 19). Furthermore, intronic peaks and poly(A) sites in non-annotated gene regions are, instead, enriched for poly(A) motifs both upstream and downstream the TES.

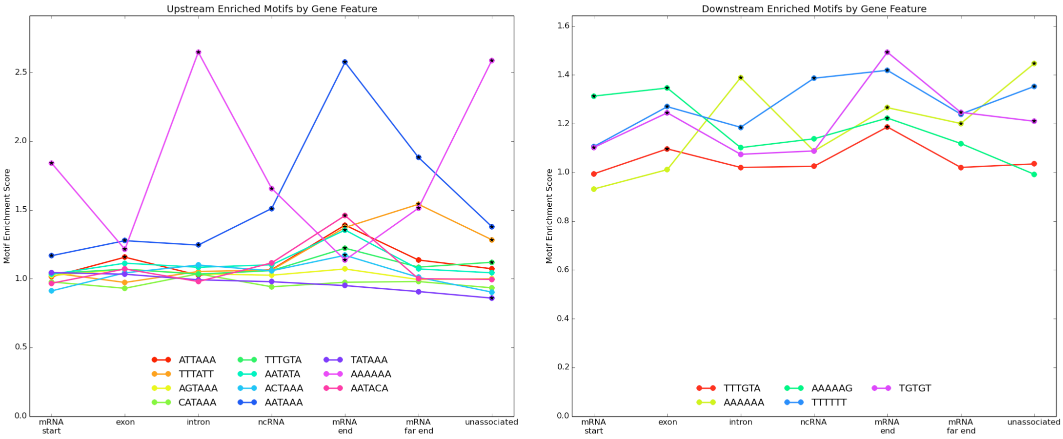


Fig. 19 Localization of polyadenylation motifs. We detected the already identified upstream and downstream polyadenylation motifs. In *D. melanoagster* they are mostly enriched at the annotated 3' end of transcripts. Some of the motifs are particularly enriched in some class of APA sites. An example is represented by the PAS variant ATTAAA enriched in exonic APAs.

Previous studies estimated the position of a number of these motifs based on a detailed analysis of a handful of examples. In this study, both the high resolution and scale of the data has facilitated a much precise analysis of the sequence and specific position of auxiliary motifs, providing an accurate map of the sequence features essential for alternative poly(A).

Conclusions and Perspectives

In this study, I have generated an accurate, comprehensive and detailed map of RNA polyadenylation at a single base resolution, for a large wild-type *D. melanogaster* population during embryonic development. To facilitate this, I first optimized a highly sensitive, quantitative and specific 3'TagSeq method that allowed us to analyze different APA usage for all genes expressed during three developmental time points. We observed that APA in *D. melanogaster* is much more pervasive than previously reported (Smibert et al. 2012a), where 65% of genes have at least two different polyadenylation sites (Fig. 17). Although approximately 10,000 genes are constantly expressed at all three time points, we observed that the vast majority are changing in the level of their expression and the usage of APA in a stage specific way (Table 1). The positional enrichment of the canonical polyadenylation signal supports the strong quality of our data: indeed we were able to *de novo* identify the major and auxiliary motifs that are bound the core polyadenylation complexes and to precisely define their position (Fig. 17 and 18).

Embryonic development is generally considered to be a highly stereotypic process since genetic studies indicate that variation in gene expression can cause catastrophic phenotype changes. As previously mentioned, this robustness of developmental reactions, in the midst of fluctuating mutations and environmental conditions within a population, is brought about through canalization (Waddington 1942), implying that variation in gene expression is buffered by different mechanisms (see 'Mechanisms of robustness'). Despite this, the process of transcription itself is rather noisy, with what appears to be leaky transcription over almost the entire genome. Changes in gene expression are major contributors to diversity in morphology, behavior, and disease among individuals, as well as to evolutionary differences between species (Emilsson et al. 2008; (Wittkopp et al. 2004; McManus et al. 2010). The majority of these have focused on adult stages or differentiated tissues (Pickrell et al. 2010), despite the central role development plays in forming phenotype from a given genotype. It remains unclear, for example, to what extent gene expression variation is tolerated during development or how the effects of genetic variants are buffered to ensure the development of stereotypically patterned embryos.

The last few decades have seen an explosion of studies aimed at connecting changes in gene expression to specific single nucleotide polymorphisms that have been called *expression Quantitative Loci* (eQTL) (reviewed in Majewski and Pastinen 2011; Rockman and Kruglyak 2006). However, there is very little information in the context of embryonic development, as no large scale study has been conducted to date.

Here, we are currently taking advantage of the quantitative 3'TagSeq data for the 80 individual isogenic lines to perform the first large-scale eQTLs analysis during metazoan development, to date. In collaboration with Ewan Birney's group, at EMBL-EBI, we are developing a method to identify and compare eQTLs that lead to variation in gene expression at all three time points or specifically in one stage. We are not only focusing our attention on SNPs but also on the effect that structural variations (SVs) could have on transcriptional regulation. Even at this early stage of the analysis, the very first preliminary results identified ~2000 expression QTLs, which surprisingly includes genes known to be critical for development (Fig. 20). We thus have a first initial indication that developmental reactions can harbor a surprising amount of functional variation affecting even essential genes in development.

Taken together, these data will provide an initial insight into the complex relationship between genotype, phenotype and developmental progression in an experimentally tractable organism. Moreover using the polyadenylation signal in combination with transcription occupancy and epigenetic data available in the lab, we will be able to investigate the molecular mechanism by which a SNP or SV leads to changes in gene expression and potentially how the regulatory elements can cope with this to ensure robustness to the developing organism.

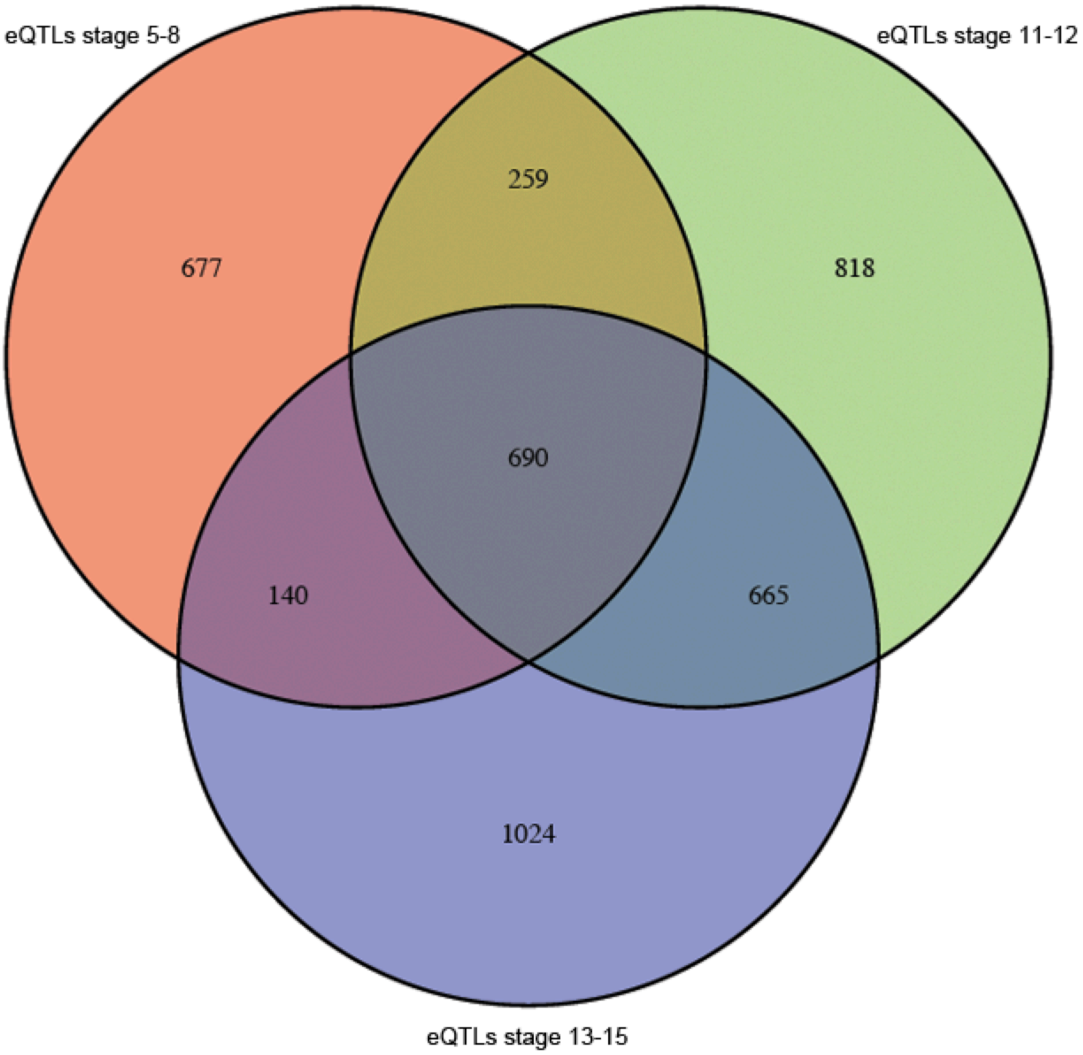


Fig 20 eQTLs in developing *D. melanogaster*. Using t-test analysis, we identified ~2000 eQTLs per developmental window. Most of the eQTLs are stage specific, while ~700 are common to all the time ponts.

References

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**: 1001–1010.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**: 53–64.

Bernstein P, Peltz SW, Ross J. 1989. The poly(A)-poly(A)-binding protein complex is a major determinant of mRNA stability in vitro. *Mol Cell Biol* **9**: 659–670.

Brown CE, Sachs AB. 1998. Poly(A) tail length control in Saccharomyces cerevisiae occurs by message-specific deadenylation. *Mol Cell Biol* **18**: 6548–6559.

Calvo O, Manley JL. 2001. Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. *Mol Cell* **7**: 1013–1023.

Castelo-Branco P, Furger A, Wollerton M, Smith C, Moreira A, Proudfoot N. 2004. Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol Cell Biol* **24**: 4174–4183.

Chen F, MacDonald CC, Wilusz J. 1995. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res* **23**: 2614–2620.

Chuvpilo S, Avots A, Berberich-Siebelt F, Glöckner J, Fischer C, Kerstan A, Escher C, Inashkina I, Hlubek F, Jankevics E, et al. 1999. Multiple NF-ATc isoforms with individual transcriptional properties are synthesized in T lymphocytes. *J Immunol* **162**: 7294–7301.

Dai W, Zhang G, Makeyev EV. 2012. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res* **40**: 787-800

Danckwardt S, Gantzert A-S, Macher-Goeppinger S, Probst HC, Gentzel M, Wilm M, Gröne H-J, Schirmacher P, Hentze MW, Kulozik AE. 2011. p38 MAPK controls prothrombin expression by regulated RNA 3' end processing. *Mol Cell* **41**: 298–310.

Danckwardt S, Kaufmann I, Gentzel M, Foerstner KU, Gantzert A-S, Gehring NH, Neu-Yilik G, Bork P, Keller W, Wilm M, et al. 2007. Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals. *EMBO J* **26**: 2658–2669.

Dantonel JC, Murthy KG, Manley JL, Tora L. 1997. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. *Nature* **389**: 399–402.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173–1183.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866.

Elkon R, Drost J, van Haaften G, Jenal M, Schrier M, Vrielink JAO, Agami R. 2012. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol* **13**: R59.

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**: 496–506.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. 2008. Genetics of gene expression and its effect on disease. *Nature* **452**: 423-428

Fong N, Bentley DL. 2001. Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev* **15**: 1783–1795.

Ford LP, Bagga PS, Wilusz J. 1997. The poly(A) tail inhibits the assembly of a 3“-to-5” exonuclease in an in vitro RNA stability system. *Mol Cell Biol* **17**: 398–406.

Gawande B, Robida MD, Rahn A, Singh R. 2006. Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline. *EMBO J* **25**: 1263–1272.

Glover-Cutter K, Kim S, Espinosa J, Bentley DL. 2008. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **15**: 71-88

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473–479.

Huang Y, Carmichael GG. 1996. Role of polyadenylation in nucleocytoplasmic transport of mRNA. *Mol Cell Biol* **16**: 1534–1542.

Huang Y, Li W, Yao X, Lin Q, Yin JW, Liang Y, Heiner M, Tian B, Hui J, Wang G. 2012. Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol Cell* **45**: 459-469

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**: 97–101.

Ji Z, Luo W, Li W, Hoque M, Pan Z, Zhao Y, Tian B. 2011. Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol* **7**:534

Ji X, Kong J, Liebhaber SA. 2011. An RNA-protein complex links enhanced nuclear 3' processing with cytoplasmic mRNA stabilization. *EMBO J* **30**: 2622–2633.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci USA* **106**: 7028–7033.

Ji Z, Tian B. 2009. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS ONE* **4**: e8419.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664–668.

Lau AG, Irier HA, Gu J, Tian D, Ku L, Liu G, Xia M, Fritsch B, Zheng JQ, Dingledine R, Xu B, Lu B, Feng Y. 2010. Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proc Natl Acad Sci USA* **107**: 15945-15950

Legendre M, Ritchie W, Lopez F, Gautheret D. 2006. Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput Biol* **2**: e43

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**: 464–469.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**: 72–79.

Mandel CR, Bai Y, Tong L. 2008. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* **65**: 1099–1122.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**: 953–956.

Mansfield KD, Keene JD. 2012. Neuron-specific ELAV/Hu proteins suppress HuR mRNA during neuronal differentiation by alternative polyadenylation. *Nucleic Acids Res* **40**: 2734-2746

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3“ end processing reveals a decisive role of human cleavage factor I in the regulation of 3” UTR length. *CellReports* **1**: 753–763.

Martinson HG. 2011. An active role for splicing in 3'-end formation. *Wiley Interdiscip Rev RNA* **2**: 459–470.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673-684

McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD, Wickens M, Bentley DL. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**: 357–361.

McCracken S, Lambermon M, Blencowe BJ. 2002. SRm160 splicing coactivator promotes transcript 3'-end cleavage. *Mol Cell Biol* **22**: 148-160

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**: 816–825.

Meinhart A, Cramer P. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* **430**: 223–226.

Meyer S, Urbanke C, Wahle E. 2002. Equilibrium Studies on the Association of the Nuclear Poly(A) Binding Protein with Poly(A) of Different Lengths †. *Biochemistry* **41**: 6082–6089.

Millevoi S, Decorsière A, Loulergue C, Iacovoni J, Bernat S, Antoniou M, Vagner S. 2009. A physical and functional link between splicing factors promotes pre-mRNA 3' end processing. *Nucleic Acids Res* **37**: 4672–4683.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773-777

Moreira A, Takagaki Y, Brackenridge S, Wollerton M, Manley JL, Proudfoot NJ. 1998. The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms. *Genes Dev* **12**: 2522–2534.

Nagaike T, Manley JL. 2011. Transcriptional activators enhance polyadenylation of mRNA precursors. *RNA Biol* **8**: 964-967

Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, et al. 2013. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**: 923–928.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**: 1018–1029.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–772.

Proudfoot N. 2004. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**: 272–278.

Proudfoot NJ. 2011. Ending the message: poly(A) signals then and now. *Genes Dev* **25**: 1770–1782.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat Rev Genet* **7**: 862–872.

Rozenblatt-Rosen O, Nagaike T, Francis JM, Kaneko S, Glatt KA, Hughes CM, LaFramboise T, Manley JL, Meyerson M. 2009. The tumor suppressor Cdc73 functionally associates with CPSF and CstF 3' mRNA processing factors. *Proc Natl Acad Sci USA* **106**: 755–760.

Sandberg R, Neilson JR, Sarma A, Sharp PS, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**: 1643-1647

Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761–772.

Shi Y. 2012. Alternative polyadenylation: new insights from global analyses. *RNA* **18**: 2105–2117.

Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB, et al. 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. *CellReports* **1**: 277–289.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**: 245-254

Takagaki Y, Seipelt RL, Peterson ML, Manley JL. 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**: 941–952.

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**: 201–212.

Tian B, Pan Z, Lee JY. 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**: 156–165.

Venkataraman K, Brown KM, Gilmartin GM. 2005. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* **19**: 1315–1327.

Vinciguerra P, Stutz F. 2004. mRNA export: an assembly line from genes to nuclear pores. *Curr Opin Cell Biol* **16**: 285–292.

Waddington C. 1942. Canalization of development and the inheritance of acquired characters. *Nature* **150**: 563–565.

Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701-718

Wahle E, Keller W. 1992. The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors. *Annu Rev Biochem* **61**: 419–440.

Wells SE, Hillner PE, Vale RD, Sachs AB. 1998. Circularization of mRNA by eukaryotic translation initiation factors. *Mol Cell* **2**: 135–140.

Wilkening S, Pelechano V, Järvelin AI, Tekkedil MM, Anders S, Benes V, Steinmetz LM. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res* **41**: e65.

Wilusz CJ, Wormington M, Peltz SW. 2001. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol* **2**: 237–246.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430**: 85–88.

Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D, Oakey RJ. 2008. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev* **22**: 1141-1146

Yao P, Potdar AA, Arif A, Ray PS, Mukhopadhyay R, Willard B, Xu Y, Yan J, Saidel GM, Fox PL. 2012. Coding Region Polyadenylation Generates a Truncated tRNA Synthetase that Counters Translation Repression. *Cell* **149**: 88-100

Yoon OK, Hsu TY, Im JH, Brem RB. 2012. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet* **8**: e1002882.

Zhang X, Virtanen A, Kleiman FE. 2010. To polyadenylate or to deadenylate: that is the question. *Cell Cycle* **9**: 4437–4449.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405–445.