Re-assembly and re-annotation of the *Xenopus tropicalis* genome for *in vivo* ChIA-PET analysis

Nicolas Buisine[1], Xiaoan Ruan[2], Patrice Bilesimo[1,5,6], Alexis Grimaldi[1], Gladys Alfama[1], Pramila Ariyaratne[2], Fabianus Mulawadi[2], Jieqi Chen[2], Wing Kim Sung[2], Edison Liu[2,3], Barbara A. Demeneix[1], Yijun Ruan[2,3,4*], Laurent M. Sachs[1*]

[1] UMR CNRS 7221, Département Régulations, Développement et Diversité Moléculaire, Muséum National d'Histoire Naturelle, CP32, 57 Rue Cuvier, 75231 Paris Cedex 05, France.

[2] The Jackson Laboratory of Genomic Medicine, 263 Farmington Ave, Connecticut, USA and The Genome Institute of Singapore, 60 Biopolis Street, Singapore.

[3] Department of Genetics and Developmental Biology, University of Connecticut, USA.

[4] Genome Institute of Singapore, 60 Biopolis Street, Singapore.

[5] Watchfrog S.A.S., 4 rue Pierre Fontaine, 91000 Evry, France.

[6] Ken and Ruth Davee Department of Neurology and Department of Cell and Molecular Biology, Northwestern University Feinberg School of Medicine Ward 10-332, 303 East Chicago Avenue, Chicago, Illinois 60611, USA.

Corresponding author: YR and LS

YR: The Jackson Laboratory of Genomic Medicine, 263 Farmington Ave, Connecticut, USA. Yijun.Ruan@jax.org

LS: "Function and action mechanisms of thyroid hormone receptors" group, UMR CNRS 7221, Département Régulations, Développement et Diversité Moléculaire, Muséum National d'Histoire Naturelle, CP32, 7 Rue Cuvier, 75231 Paris Cedex 05, France. Phone: 33 (0)1 4079 3604, email: sachs@mnhn.fr

Keywords: Thyroid hormone signaling, ChIA-PET, Paired end technology, genome assembly, genome annotation.

Abstract

Genome-wide functional analyses require high-resolution genome assembly and annotation. We applied ChIA-PET to analyze gene regulatory networks, including 3D chromosome interactions, underlying thyroid hormone (TH) signaling in the frog *Xenopus tropicalis*. However, the available versions of *Xenopus tropicalis* assembly and annotation did not have the resolution required for ChIA-PET, nor for physiological analysis. To overcome these limitations, the current genome assembly and annotations were improved using the paired end tag (PET) sequencing technologies and approaches (e.g., DNA-PET [gPET], RNA-PET and ChIA-PET etc.). The large insert (10 kb, 15 kb) paired end DNA-PET with high throughput NGS sequencing not only significantly improved genome assembly quality, but also strongly reduced genome “fragmentation”, reducing total scaffold numbers by ~60%. Next, RNA-PET technology, originally designed and developed for full length coding and fusion mRNA in whole transcriptome studies (ENCODE consortia), was applied to capture full length transcripts specifically at the 5' and 3' ends, thereby defining TSS and TTS boundary sites and providing more accurate gene annotations at the genome-wide level. These amendments in assembly and annotation were essential prerequisites for the ChIA-PET analysis, enabling identification of complex regulatory configurations around TH-regulated genes and putative networks underlying physiological responses. This demonstration of the application of the ChIA-PET technology to understand physiological genetic regulations in a less-conventional model could provide valuable conceptual and methodological guidance for similar approaches on other species with low-resolution genome data.

Introduction

The rapid advances and reduced costs of sequencing are increasing the number of genomes available for study (Mardis, 2011). However, functional post-genomic analyses require high quality genome assembly and annotation, and such levels have only been achieved for a few reference genomes including human, mouse, drosophila and yeast. Initial genome assembly and annotation is labor-intensive, implicating sustained computation and manual curation efforts that will most often be incomplete. Regular curation and updating are required to ensure high standards in reference genome quality in current public databases (Reese et al., 2000, The ENCODE Project Consortium, Graveley et al., 2011, Hoskins et al., 2007, Lin et al., 2007, The modENCODE Consortium). In contrast, most newly completed genome projects targeting *de novo* assembly, produce draft genome assemblies and annotations with tens, or even hundreds of thousands of fragmented scaffolds, numerous gaps and sequencing errors (see for example Varshney et al., 2013, Bombarely et al., 2012, Li et al., 2009). Although these assemblies, together with genome annotation, provide some estimation of the gene repertoire, they carry little information on the medium/large scale structure of the genome. Too often, the quality of draft genomes is a major limiting when trying to exploit maximally data from recent advances in functional genomics. Alternative strategies and approaches are needed to improve assembly and annotation (Huang et al., 2013, Eckalbar et al., 2013, Zhuo et al., 2012 for a few examples). Current advances in high throughput sequencing of genomic DNA and RNAs offer complementary technologies to re-annotate genomes and to use these high-resolution tools to shed new light on our perception of physiological and biological processes in non conventional models.

In all chordates, post-embryonic development involves a complex set of molecular, cellular and anatomical transitions, which are orchestrated by thyroid hormones (TH) signaling (Tata, 2006, Laudet, 2011). The anatomical changes are often dramatic, as it is well illustrated in the case of flatfish and amphibian metamorphosis. In precocial birds, such as chicks, this transition corresponds to the hatching period. In mammals it is often equated to the perinatal period, for which the transition phenotype is anatomically subtler, but the changes in respiration, nutrition and sensory nervous system maturation are well marked. THs triggers complex transcriptional remodeling programs affecting cell fate by inducing a variety of transcriptional programs in different tissues/cell types, for instance apoptosis/cell death versus cell growth/proliferation. These programs often occur simultaneously in different cells within given organs or tissues. Thus, because of these cell specific responses, it is difficult to design experimental approaches to characterize the transcriptional programs underlying the responses, particularly in *in vivo* settings. To tackle this biological question and challenge, we use metamorphosis in *Xenopus (Silurana) tropicalis* (*X. tropicalis*) as a working model. Amphibian metamorphosis is the most marked of the developmentally TH-dependent post-embryonic transitions described above. During anuran (frogs and toads) metamorphosis an aquatic tadpole changes to an air-breathing frog, with almost all tissues undergoing profound remodeling (Shi, 1999, Tata, 1993). A key feature of amphibian metamorphosis is that external organs can display opposite fates: the tadpole tail regresses and ultimately disappears whilst limbs grow *de novo* from limb buds. Amphibian metamorphosis is thus an attractive model to study the impacts of THs on transcriptional reprograming during post-embryonic development in vertebrates. As in other organisms, THs signaling is mediated by nuclear receptors (TR) affecting the transcriptional state of target genes (Grimaldi et al., 2013). Our overall aim is to determine the repertoire of TR mediated direct and indirect TH target genes and the characterization of TR-based regulatory networks.

Deriving a full repertoire of a transcription factor binding sites (TFBS), such as TR, can usually be achieved with a target-specific ChIP-Seq experiment, in which the DNA sequences bound to an affinity-purified transcription factor are sequenced (Furey, 2012). The problem of genome wide profile maps is that TFBS are often located far away (>200 kb) from their cognate gene, and can even be located within another gene. Indeed, enhancers located far from their target genes functionally and physically interact with the transcriptional machinery at the target gene through DNA looping (Sanyal et al. 2012). As a result, it is misleading to relate each TFBS to a given target gene solely based on genomic proximity. In addition, TFBS can also be arranged into *cis* regulatory modules controlling the transcriptional status of several target genes at once, which may further complicate the assignment of a

TFBS to its target genes (Li et al. 2012). In order to circumvent these problems, we used ChIA-PET technology to identify TR binding sites across the whole genome in a manner similar to obtaining ChIP-Seq binding site information, but also revealing their physical interactions with distant genomic sites (Fullwood et al. 2009a, Fullwood et al. 2009b). This approach ensures identification of the gene sets directly regulated by TR-mediated signaling. Crucially, successful ChIA-PET analysis relies on high quality genome assembly and annotations, as “fragmented” assemblies lower the resolution of large-scale interaction maps. In addition, poor gene annotation, especially at 5' ends, would tend to associate transcriptional regulators to regions located outside of genes and limit the identification of direct target genes. Unfortunately, during the initial course of the ChIA-PET analysis of the TR transcriptional network for *X. tropicalis* metamorphosis, it became clear that the state of the genome assembly and annotation was the bottleneck of the analysis, as the resolution was too low to resolve accurate ChIA-PET locus assignment.

Here, we describe the technological framework used to improve annotation and re-scaffold the genome assembly of *X. tropicalis*. This genome is also an ideal test case because the genome size (~1.5 Gb) is smaller than mammalian genomes, and the draft assembly, although “fragmented” (Hellsten et al. 2010, ~ 20,000 scaffolds) is much less so than others (e.g. fish *Platipus lavaretus* with ~350,000 scaffolds, Warren et al., 2008, or, Panda *Ailuropoda melanoleuca* with >500,000 scaffolds, Li et al., 2010). We used large insert DNA-PET and full-length RNA-PET (in a combination with paired end RNA-Seq) to define boundaries of genome-wide transcripts using NGS sequencing. This combined approach allows accurate capture of 5' and 3' transcript ends and variants, in a length-independent manner. The method(s) are remarkably powerful for improving gene annotation, permitting detection of transcripts and unambiguously identifying gene boundaries, alternative transcripts, transcript start sites (TSS) and termination sites (TTS) etc. We also used the connectivity information provided by the large insert DNA-PET to bridge “un-assembled scaffolds” in the published assembly. This combined effort dramatically reduced the total scaffolds in the genome assembly, corrected and improved the published version and enabled ChIA-PET analysis.

RESULTS

Large insert DNA-PET significantly reduced complexity of genome assembly.

Our starting point was the published *X. tropicalis* genome assembly (v4.1). It is composed of 19,759 scaffolds, with the top ~ 2000 scaffolds (longer than 20kb) covering 80% of the assembly (~ 1.5Gb) (Hellsten et al. 2010). The average scaffolds size is about 76kb, which is smaller than the average gene length (~ 28kb) and would be a limitation if one wants to address physical interactions across several genes (Buisine and Sachs, 2009). The vast majority of these scaffolds are not fully sequenced, leaving numerous, often large, assembly gaps (Figure 1A). The scaffolds can be classified into three main groups: 1) numerous small "scaffolds", composing ~30% of the total scaffolds, ~1.4% of the total sequence, actually have the size of Sanger sequencing reads or short contigs (up to 5kb); 2) a collection of relative larger scaffolds (15kb to 50kb) have average sizes consistent with that obtained from cosmid clones. Given that they comprise large assembly gaps, they are more likely derived from sequences originating from cosmid inserts. Finally, 3) the largest collection of scaffolds (> 100kb), that are most likely assembled from BAC clone sequences. The quality of this assembly is very similar to that of other draft assemblies (supplementary data in Buisine and Sachs, 2009). Of note, among the 183,010 assembly gaps present in the current published sequence, 84126 (45.9%) are exactly 50bp long, which probably reflects a technical issue of the assembly process. Such systematic bias suggests that the real size of these "50bp gaps" is unknown and was set by default at this value.

DNA-PET technology was used to improve this assembly. This technology was initially applied to cancer research (Ruan et al., 2007) and more recently genome assembly (Yao et al. 2012). The method sequences both ends of size-defined large (e.g., 10kb or 15kb) DNA fragments, which are then mapped to the reference genome (Supplementary Figure 1). The approach provides important connectivity information over large genomic regions, which is often critical for genome assembly. The sequenced reads composed of two kinds of PETs: "concordant" PETs (cPETs), for which the two tags are mapped onto the same scaffold at the given distance corresponding to the size of the DNA fragments; and "discordant" PETs (dPETs), for which the two tags mapped onto different scaffolds. cPETs can be used to assess the quality of the datasets, estimate the number of assembly errors, the real size of assembled gaps of unknown length ("50bp gaps", see below) and possible structural variations. dPETs can also be used to improve the assembly of the scaffolding. Two large insert DNA-PET libraries (10kb and 17kb) were created using kidney and liver genomic DNA. A summary of the DNA-PET statistical data is shown in Figure 1B. A total of 77.8 million uniquely mapped PETs were obtained for further mapping and downstream analysis.

The average insert size (mapped distance between two tags), estimated by measuring the mean genome span of cPETs, was 9.6kb and 17.5kb for the two libraries IXT010 and IXT011, respectively. This agrees well with the expected size (Supplementary Figure 2A) and validates library quality. As expected, concordant PETs composed the vast majority of the dataset and dPETs accounted for 8 to 10 % of the total PETs (Figure 1B), consistent with what is usually found in most DNA-PET datasets. In addition, genome-wide cPET coverage profiles were computed from each library by counting the number of cPETs in sliding 20kb windows. They are highly correlated (*r*=0.97, Supplementary Figure 2B), thus showing the high quality and reproducibility of the datasets.

We first used the cPET datasets to estimate the size of the numerous "50bp gaps". To this end, we selected the cPETs spanning a single gap of exactly 50bp and computed the average difference between the expected (9.6 or 17.5kb) and the observed genome span of individual PETs. Positive values of this estimate indicate that the minimal region delineated by the cPETs is actually much longer than expected, presumably (but not only) because of the uncertainty attributed to the gap length. Negative values, meaning that the minimal region is actually shorter than expected, may result from structural polymorphism between our sequenced strain and the reference genome sequence. We found that more than half (46,249) of the "50bp gaps" out of (84,126) have an actual average size of ~500bp long (Supplementary Figure 3). This increases the total size of the assembly by an extra 12.7Mb. In a typical case, the genome span of cPETs follows a bell shape distribution (see Supplementary Figure 4 for a few illustrative examples). Interestingly, we also spotted a number of cases (~ 8%) where the genome span distribution of cPETs was bimodal, of which one, or both gap length estimates deviated significantly from expectation (Supplementary Figure 5). A similar procedure was used for assembly gaps of other sizes, of which a few examples of mono and bi-modal distributions are shown (Supplementary Figures 6 and 7). Altogether, the analysis increased the assembly size together by ~ 21.8Mb in total. Simultaneously, our analysis also re-calculated the size of assembly gaps, since this can have a significant impact in estimates of the genomic distances between structural and/or functional elements. Furthermore, 50bp long features can be difficult to see on genome browsers and may be visually missed by biologists.

In a number of cases, the cPET coverage drops to zero, meaning that the left and right hand sides of the breakpoint are not connected together at the resolution of the average DNA-PET insert size (~10kb and ~17kb). These assembly breakpoints can either result from structural polymorphisms between reference and sampled genomic DNA, or from various *scenarii* of mis-assembly errors (*e.g.* the left and right hand sides of the breakpoints are each connected to a different scaffold, or the "50bp assembly gap" actually correspond to a region much longer than ~20kb). Two examples are illustrated in Supplementary Figure 8. In both cases, cPET coverage drops to zero at a 50bp assembly gap. Note that by definition, the length of these 50bp gaps cannot be estimated based on the average cPET span. Interestingly, the breakpoints are flanked by one or two clusters of discordant PETs pointing to different scaffolds, strongly suggesting that based on our sequenced DNA sample, the left and right sides of the breakpoints are each connected to a different scaffold. Overall, we identified a total of 352 cPET coverage breakpoints, of which, 211 were re-connected to other scaffolds during the re-scaffolding process (see below).

We then used dPETs to improve the physical contiguity of the scaffolds and to better describe the large-scale structure of the *X. tropicalis* genome. As a first step, and given the number of large assembly gaps that are susceptible to accommodate smaller scaffolds, scaffolds were split at gaps larger than 10kb. This corresponded to 1031 locations. We also split scaffolds at cPETs coverage breakpoints (see above). The two datasets were then combined and processed with the PE-Assembler software (unpublished method), which was previously shown to perform well (see methods, Earl et al. 2011). The statistics of the improved assembly are shown in Figure 1C. The total number of scaffolds is reduced by 56%, the N50 and the size of the largest scaffold increases by ~3 folds (2.8 x and 3.5 x, respectively). A total of 4901 short contigs (~ 2-3kb "scaffolds") could not be connected to other scaffolds. These originate from small contigs which are too short to be assembled because given the insert size of our DNA-PET libraries (> 9.6kb), two contiguous short scaffolds may not be directly connected by dPETs and would fail to assemble into longer chains. The remaining scaffolds could be associated into longer chains, of which 1283 are composed of 5 scaffolds or more.

In order to visualize the improvements, we plotted for each scaffold the relationship between its size in bp (in log scale) and the percent of its sequence remaining to be determined (Figure 1D) and compared the result with the same representation of the published assembly (Figure 1A). Scaffolds located in the uppermost part of the plot correspond to scaffolds containing numerous or large assembly gaps. In contrast, scaffolds located at the lowermost part of the plot contain few, if any. Ideally, assembled scaffolds sequenced to completion, and ultimately chromosomes, should be located at the lower right side of the plot. Re-scaffolding with DNA-PET clearly shifts the distribution of the scaffolds corresponding to 80% of the sequence assembly (pink rectangle, Figure 1A) from 1Mb to towards to right hand side of the plot (Figure 1D).

Re-scaffolding with dPETs connected 26,703 scaffolds into 3594 "chains" of various length, of which the frequency distribution is shown Figure 1E. The longest chain is composed of 519 scaffolds with an average chain length of 3.14 (median = 1). Although this may suggest that the re-scaffolding efficiency is rather low, finally, only a minority of scaffolds (4901, ~ 18%) were not linked to another scaffold. When discarding these short "scaffolds", the average chain length raises up to 5.12 (median = 3). In addition, smaller scaffolds filled 835 assembly gaps. An example of re-scaffolding is shown Figure 1F. The two libraries, built with different insert sizes, provided similar results, although the PET count of the edges originating from the 17kb library tends to be higher than that from that of the 9.6kb library. Two scaffolds containing large assembly gaps (105kb in scaffolds_481 and 5 kb in scaffold_409), which were split in two before re-scaffolding, were effectively re-connected. Overall, the estimated link size between two scaffolds varies from < 1 to ~ 10kb. In a number of cases, with link size < 6 kb, we could validate the link by conventional long-range PCR (data not shown).

Connectivity between scaffolds can be represented by direct acyclic graphs (Supplementary Figures 9-11), where vertices correspond to scaffolds (drawn as ovals) and dPET links to edges. In these figures, vertices are labeled with the scaffold name, size and its original position in the scaffold, before the splitting at assembly gaps and assembly errors. With this representation, bubble-like structures correspond to small scaffolds inserted into larger ones. They can also result from the complex the connectivity between tandemly arranged short scaffolds, or a combination thereof. Note that bubble-like structures are common but bifurcating forks are rare, meaning that there are only a few unresolved connections. This result further illustrates the strength of our dataset. A number of long scaffold chains, long-range connectivity and bubble-like structures are shown in Supplementary Figures 9, 10 and 11. Detailed examples of assembly gaps filled-in by small scaffolds are shown in Supplementary Figures 12, 13 and 14.

In summary, in this section we show that DNA-PET helps improve genome assembly by constructing large chains composed of connected scaffolds that enable one to identify and correct assembly errors. From here on, our results are based on the assembly v4.1 after re-scaffolding.

Genome re-annotation with RNA-PET and RNA-Seq.

Gene annotation is a complex process relying on high quality experimental resources and dedicated computational tools. The transcriptional start site is an important functional feature, which is usually difficult to annotate, especially with conventional RNA-Seq, or EST mapping. Here, we used a combination of paired end RNA-Seq (PE-RNA-Seq), conventional RNA-Seq and RNA-PET datasets to re-annotate the *X. tropicalis* gene content.

PE-RNA-Seq based transcript assembly and mapping were carried out following a standard procedure with TOPHAT (Trapnell et al., 2009) and CUFFLINKS (Trapnell et al., 2010). The resulting models were then combined with RNA-PET data. RNA–PET was initially a technology designed for detecting transcripts resulting from gene fusion in cancer cells (Ruan et al., 2007) and recently used to derive a *de novo* annotation of the sweet orange genome (Xu et al., 2012). RNA-PET demarcates specifically both 5’ and 3’ ends (27bp each) of all expressed full length RNA molecules, which include normal and splice variants, truncated isoforms, fusion transcripts etc. that might be derived from various conditions. The protocol used is a modified version (Ruan & Ruan, 2012) of the original method described as “GIS-PET” (Ruan et al., 2007). This process is summarized in supplementary Figure 15.

A total of seven RNA-PET libraries were constructed using RNAs isolated from two larval tissues (i.e. premetamorphic stage NF54 tadpole tailfin and limb buds), and five adult tissues (i.e. brain, kidney, liver, muscle and intestine). After paired end sequencing at 2 x 36bp, the 3' end tag of the transcript, specifically recognized by a signature sequence "AACTGCTG" (Ruan and Ruan, 2012), was identified through a modified Smith and Waterman (SW) algorithm. Practically, each paired end sequence read (2 x 36bp) was screened first for the 3’-specific “signature” sequence, and labeled as HT (Head and Tail) if one end had found the “signature”, labeled TT (Tail Tail) if both ends had the “signature” (noise), or HH (Head Head) if neither end had the “signature” (noise). For each library, between 60% to 75% of PET sequences had HT sequences, indicating that the majority of the captured PETs represent full-length transcripts (Supplementary Figure 16). Approximately 20% to 40% PET sequences are noisy HH products largely due to sequencing errors in the ‘T’ signature. We noted, however, that relaxing the SW algorithm parameters (*i.e*. lowering the alignment mismatch threshold) failed to rescue more HT PETs from the HH pool. Thus, this minor artifact is a direct consequence of searching a short (8bp) signature sequence, which cannot tolerate more than two bp mismatches. Subsequent analysis was carried out only using the desired HT (5’- and 3’-tag) sequence reads.

All libraries provided quality results, except the liver library that produced a limited number of PETs, which probably resulted from a library of lower quality. In order to address possible sequencing bias, we also built three libraries (from intestine, kidney, liver) for sequencing on a SOLiD platform. Note that due to the SOLiD sequencing chemistry limits, the 5' and the 3' end of transcripts cannot be distinguished and thus this data can only provide transcript boundaries, irrespective of the transcription orientation. However, the coverage of the transcription units with transcripts models between the two sequencing platforms was highly correlated *r*=0.867, illustrating that the RNA-PET based transcript models have little bias from different sequencing platforms.

PET clustering was carried out independently for each library with two sets of parameters (see Methods). The set of parameters is quite stringent and is less sensitive to noise at the cost of producing shorter clusters. The relaxed set of parameters tends to produce longer models, but may merge together a few overlapping clusters. The two resulting datasets were then merged and filtered out by selecting independently for each transcription unit the longest model without over-clustering. This "promotion" step adapts the stringency of the clustering parameters defined by the local context. Clusters were subsequently coalesced in order to derive gene new models.

In order to characterize the noise originating from unspecific ligation events, we plotted for each library the frequency of each cluster as a function of its size (in bp, with a logarithmic scale, see Supplementary Figures 17-23). It is reasonable to assume that background noise would be associated with clusters with a low PET count, whereas PET count would be more robust for clusters derived from highly expressed genes in the tissue. In all PET libraries, the frequency of cluster size followed a bimodal distribution, with a sharp drop at 3.3 (~ 2Kb). When only clusters with increasing PET count were taken into account, the first peak progressively disappeared, indicating that cluster span smaller than 2kb were associated with low PET count noise and were discarded. Putative transcript models were further filtered out if they were not supported by at least 2 reads per kb, based on strand-specific conventional RNA-Seq carried out from the same RNA samples (see Methods). On average, PETs could be clustered into 712,817 clusters (Supplementary Figure 24) that may correspond to as many alternative transcripts. Putative transcript models were further coalesced and were assigned as 17,993 distinct genes (or transcripts).

Overall, the RNA-PET and RNA-Seq identified 12,387 expressed genes already annotated in the Ensembl database from the tissues tested. A total of 4,729 genes described in Ensembl were not expressed, probably representing exclusion from these tissues at the developmental stages examined.. Moreover, we also identified 5,607 previously un-annotated transcription units. Approximate 80% of the RNA-Seq reads were located within the boundaries of the new models (Figure 2A). In contrast, around ~ 63% of RNA-Seq reads overlap with Ensembl gene models, indicating that improvement in description/annotation is required for the existing assembly on gene and transcription units. Moreover, Ensembl models specifically capture only a marginal fraction of RNA-Seq reads, suggesting that only a fraction of transcribed units may be missed or that missed genes have low expression levels. Altogether, these results showed that gene re-annotation, based on the tissue RNA sequenced, significantly improved transcript annotation. In order to further characterize the improvement of gene annotation, we plotted the density of RNA-Seq reads and the RNA-PET coverage over the normalized length of Ensembl gene models, together with 10kb upstream and downstream (see Methods, Supplementary Figure 25). RNA-Seq reads are mostly found within gene bodies, although a significant fraction of them extends beyond the gene boundaries into the 5' and 3' flanking regions. The reciprocal plot showing that the new annotation encompasses most Ensembl genes and efficiently captured RNA-Seq reads confirms this result.

Moreover, our new models are significantly longer than Ensembl gene models (average 34,306bp vs 27,688bp). We found 58% of Ensembl genes to be significantly longer (> x 1.3 or > 5kb) once re-annotated. In order to further illustrate this point, we plotted individual gene length based on the Ensembl versus the improved annotations (Figure 2B, Supplementary Figure 26). Importantly, it was noted that RNA-PET could also capture the 5' end of 938 genes, which were missed with PE-RNA-Seq, which further illustrates the value of RNA-PET for precisely demarcate gene boundaries. Overall, our RNA-PET data detected a total of 187,396 alternative TSS. A few illustrative examples of improvement are shown in Figure 3 (A to C). The cadm2 gene is an extreme case (3B), for which the Ensembl gene model is 210kb long, but the RNA-PET model extends the gene boundaries by an extra ~730kb. The resulting new gene model, strongly supported by RNA-Seq data, is ~940kb long. This result illustrates the value of RNA-PET in improving annotation by accurately demarcating gene boundaries. Gene size can indeed vary over several order of magnitudes whereas transcript size, which is ultimately measured by RNA-PET, is much shorter, more homogeneous and thus, easier to capture. To confirm that RNA-PET efficiently captures transcripts ends, we relied on the fact that the 5' and 3' ends of transcribed genes is RNA Pol II enriched (Adelman and Lis, 2012) and carried out a chromatin immuno-precipitation (ChIP) with an anti-RNA Pol-II antibody. RNA Pol-II ChIp was followed by deep sequencing (ChIP-Seq), and derivation of genome-wide density profiles. As shown in Figure 2C, RNA Pol II is strongly enriched at 5' and 3' ends of RNA-PET models. Furthermore, the RNA-Pol II peak at the 5' end of genes is 31bp offset relative to the transcriptional start site defined by RNA-PET based models (Figure 2C). This result is consistent with previous observations where genome wide profiling showed that RNA Pol II density peaks between 25 to 45bp downstream from the TSS (Zhou *et al.* 2012). In the case of the Ensembl gene models, the RNA Pol II profiles are also enriched at their 5' and 3' ends, but to a much lesser extent and the offset, relative to the gene model start, is much lower (< 10bp, Supplementary Figure 27). Given that RNA Pol II elongation pauses at a checkpoint located ~40bp downstream of the TSS, these results strongly support the view that RNA-PET helps capture the 5' end of genes more precisely and thus helps to improve gene annotation.

Interestingly, an additional ~ 800 RNA-PET models, well supported by RNA-Seq, were found to overlap two scaffolds, reflecting the fragmented state of the assembly. In most cases (~ 80%), however, the two scaffolds actually belonged to the same hyper-scaffold, as they were linked together by clusters of discordant DNA-PETs (Figure 3D). These examples cross-validated the DNA-PET and RNA-PET dataset consistency.

Altogether, these results clearly showed that 1). RNA-PET based gene models are strongly supported by RNA-Seq and RNA Pol II ChIP-Seq data, and 2). Existing gene models are extensively improved.

Benefit of DNA-PET and RNA-PET for ChIA-PET analysis.

Our overall aim at the outset of this work was to apply ChIA-PET analysis of TR binding to increase knowledge of the gene networks controlling metamorphosis. The details of the experimental procedure and the full datasets generated by this method applied to *X. tropicalis* will be published elsewhere. Here we selected salient statistics and examples to illustrate the benefit of the re-annotation process facilitated by RNA-PET, combined with the deeper analysis achieved by DNA-PET.

An initial analysis based on Ensembl models and our preliminary ChIA-PET datasets, revealed that the 5' ends of 21 genes displayed long-range interactions. Applying our improved annotation significantly improved the resolution of this ChIA-PET analysis, increasing the number of genes with long range interactions to 175. DNA-PET had a lower impact; with the rescue of 31 TR binding sites engaging discordant long-range interactions (*i.e.* connected to a functional element located on an other scaffold).

The benefit of improving genome annotation with RNA-PET, *per se*, is best illustrated with the *bcl6* gene, the boundaries of which have been significantly extended with RNA-PET (Figure 4). This gene is strongly induced by T_3 (~ 4 fold) and there is a TR binding site (BS1) located nearby (~ 65kb upstream, Figure 4). Three other TR binding sites (BS2, BS3 and BS4) are located much further downstream, 152kb and 485kb (BS2 and BS3 are separated by ~600 bp). The binding sites BS2 and BS3 are located at the 3' end of a gene newly annotated by RNA-PET (not regulated by T_3). Thus, based on the Ensembl annotation and ChIP-Seq-like data alone (e.g. the ChIP-Seq component of ChIA-PET), one would infer that *bcl6* transcription would be under the control of the TR bound at BS1. Interestingly, the binding sites BS2, BS3 and BS4 are connected by interaction PETs, suggesting that they physically interact with each other through DNA looping (Figure 4A). They are also connected to the 5' end of the new *bcl6* gene model, but not the Ensembl gene model. The site BS1 does not engage long-range interactions. Therefore, with this information one infers that bcl6 transcription would be under the control of the TR bound at BS2, BS3 and BS4, and not BS1 as initially assumed. This result is confirmed by conventional ChIP-qPCR where TR binding at BS2, BS3 and BS4 is strongly induced upon induction with T_3, but not BS1 (Figure 4B and 4C). TR binding is also enriched at the newly annotated *bcl6* TSS, to a weaker but significant (p=0.044) extent. In addition to *bcl6*, the transcripts level of the *lpp* gene is also significantly increased upon treatment (p=0.044, Figure 4D). Conversely, the transcription of *trpg1* is not affected by T_3 treatment. A model of *bcl6* locus long-range interactions through DNA looping is shown in Figure 4E.

Another example of the crucial input from RNA-PET re-annotation for ChIA-PET analysis is shown in Figure 5. A novel gene was detected by RNA-PET, this gene not showing up in the Ensembl track (Figure 5A, top three tracks). The transcription of this novel gene is strongly induced upon T_3 treatment (~ 50 fold, Figure 5A, bottom tracks). ChIA-PET analysis revealed that a strong TR binding site is located precisely at the TSS. These results were further confirmed by RT-qPCR and conventional ChIP-qCR (Figure 5B and 5C).

Altogether, these results illustrated that accurate gene annotation and re-scaffolding are critical prerequisites for ChIA-PET analysis. This may not be an issue for genome analysis of model species, but the lack of accuracy of an assembly is a severe limitation for post-genome functional studies in non-conventional models. In this respect, combinational analysis with RNA-PET and DNA-PET proved to be a reasonably efficient way to improve gene annotation and genome assembly.

Discussion

To better understand the molecular mechanisms controlling onset of specific transcriptional programs during metamorphosis we carried out an *in vivo* ChIA-PET analysis of TR binding sites in the *X. tropicalis* genome*.* ChIA-PET offers a key advantage over conventional ChIP-Seq analysis in that it simultaneously provides the equivalent of both ChIP-Seq plus 3C/4C analyses together. It thus accelerates and optimizes identification of direct target genes. However, ChIA-PET and conventional functional genomic technologies require high-resolution genome annotation. Given that a sufficiently precise annotation of the *X. tropicalis* was not available we had to improve genome assembly and gene annotation to exploit ChIA-PET analysis. The principal finding emerging from this work is the benefit of using PET technologies for improvement of genomic resources for non-conventional models.

PET technology: a powerful way to advance genome assembly and annotation

Sequencing technologies have dramatically improved over the past few years, increasing sequencing depth and reducing cost. The benefits of the PET technology, in improving genome assembly and annotation, are not limited to ChIA-PET analysis, but also extend to many aspects of genome analysis for which the quality of genome assembly and annotation have proved critical for genomic and post-genomic research. PET technology overcomes some of the difficulties encountered when improving existing assemblies and annotations (Yandel and Daniel, 2012). Although many tools exist, it usually requires to re-run complex annotation and assembly pipelines. PET technology provides a quick and easy way around this since the PET data can be used in a simple post-processing step to extend existing gene models and does not require extensive human and bioinformatic resources. A similar pipeline can be applied to the many existing (and future) draft genome assemblies for which gene annotations can be quickly improved. These technologies can also be included in the analysis pipelines used to generate *de novo* genome assemblies and gene/transcripts annotations. The description of genomic features can thus improve quickly and will benefit in turn from full power of high-resolution functional genomics, in describing the molecular mechanisms of phenotypic, genetic and epigenetic diversity. Scientists working on conventional models may also be interested in monitoring the level of genomic divergence between the reference genome sequence and that of the cell lines/strains/individuals they are working with. They might also want to annotate the transcripts of specific tissues or cell types poorly represented in databases. Importantly, DNA-PET and RNA-PET is fairly cost-effective and do not require the combined workforce of several research institutes. This could be instrumental to bring genome projects within reach of small research communities.

X. tropicalis **genome assembly**

DNA-PET allowed us to improve the published *X. tropicalis* genome assembly, reducting fragmentation by half, allowing estimation of gap size and the correction of several mis-assemblies. During the course of this work, a new version of the assembly was released (v7.1), with the inclusion of a *X. tropicalis* genetic map. This assembly reaches chromosome-level scaffolding (Wells et al., 2011). We note, however, that although 80 % of the assembly corresponds to only 8 scaffolds, the total number of scaffolds remains large (7730) with many small scaffolds (1518 between 10 and 200kb, and 6039 < 10kb), suggesting that many smaller-scale connections are left unresolved. Indeed, a large number of assembly gaps are exactly 100bp (17933 out of 44685, 40%) or 10kb (1200, ~3%) and may integrate smaller scaffolds in a manner very similar to the results presented above. In addition, based on our DNA-PET data, we found numerous (6479) assembly breakpoints together with a large number of genomic inversions (1845) of various sizes (3 to 74kb, median 16kb, with a vast majority at 10, 20 or 40kb). This illustrates that despite an apparent better assembly at the chromosome level, this assembly suffers from numerous smaller scale errors that can be easily identified and corrected with our DNA-PET datasets. Given that this level of fragmentation (breakpoints + inversions) surpasses that of version 4.1, we based our work on version 4.1.

The improvement of genome assembly, by increasing the average scaffold size up to ~ 4Mb with DNA-PET, proved less critical in our case to address ChIA-PET analysis, since the total number of rescued long range interactions is fairly low. Although this may not be true for all genomes/DNA-bound proteins, physical connections among genes and regulatory elements tend to occur in genomic domains spanning tens of kb up to a few Mb (Gibcus and Dekker, 2013). Very recent work suggests that this may indeed be a general feature of transcription factors connectivity networks (Sanyal et al, 2012). Thus, the fact that DNA-PET rescues a limited number of long-range interactions between functional elements may rather reflect an intrinsic property of the TR connectivity network instead of a failure of DNA-PET to rescue existing long-range interactions.

Nevertheless, in the near future, the data provided here when integrated with versions 4.1 and 7.1 of the genome will allow the Xenopus community to release a more comprehensive assembly.

X. tropicalis genome annotation

We used an original method to improve the gene annotation of the *X. tropicalis* genome. Indeed, RNA-PET was initially designed (Ruan et al. 2007) to discover novel fusion transcripts in cancer cells and used in combination with DNA-PET analysis to detect various genome rearrangements and structural variations (such as bicistronic transcripts, transplicing transcripts, translocation generated transcripts, deletion, insertion, tandem replication-derived transcripts). This application relies on the assumption that gene annotation is accurate and comprehensive, which is the case for human and mouse genomes. Here, we applied it in a reverse manner, to re-annotate a genome and re-define gene models, and allowing detection of novel genes.

By capturing the 5' and 3' end of full-length transcripts, RNA-PET provides direct evidence for locating TSSs and TTSs. This feature was instrumental for re-defining boundaries of gene models. Indeed, transcribed genomic regions (exons) are easily identified with conventional RNA-Seq. However, without additional information, it is difficult to connect (or not) intergenic RNA-Seq signal to nearby gene. This is well illustrated in the case of a transcribed region located just upstream of a given gene model. Without additional information, it is difficult to tell whether this is a novel gene or whether this corresponds to a first exon, or that the nearby gene model should be extended. By delineating transcript boundaries, RNA-PET unambiguously solves this issue and defines cognate promoter regions. This point is crucial since most functional genomic studies focus on transcription regulation, which relies heavily on accurate and comprehensive annotation of TSSs and flanking sequences (promoter regions). Indeed, we found that the re-definition of gene boundaries with RNA-PET is a prerequisite for our analysis of ChIA-PET data. The ChIA-PET thus permits quantitative definition of physical interactions of TR with the basal transcription machinery which can show up as direct interactions between the regulator and RNA Pol II bound at the 5' end of genes, thus unambiguously identifying direct target genes. Importantly, as these physical and remote interaction sites can span over large genomic distances (> 200kb), our results illustrate the exquisite resolution capacity of ChIA-PET in identifying transcriptional regulators even when spread over large genomic regions and at large distances from their target gene. By clearly identifying the TSS and promoter regions, this new annotation resource may also be highly relevant for people working on transcriptional regulation in *Xenopus*.

Of note, although RNA-PET does not provide information relative to the internal structure of transcripts, it nonetheless helps discriminate between alternative transcripts, which can share their 5' (or 3') end but not their 3' (or 5') end. Transcript reconstruction with paired-end RNA-Seq data helps document their internal structure, but is not guaranteed to capture the 5' end of transcripts. Paired end RNA-Seq has nonetheless the advantage of being less demanding in terms of sample quantity and library construction. Thus, even when extensive RNA-Seq (conventional or paired-end followed by transcript assembly) data are available, RNA-PET is a precious complement. In addition to improve the definition of annotated genes, we could also detect numerous RNA-PET models (5' and 3' ends of transcripts) split over two scaffolds, illustrating further benefit of using RNA-PET to correct and improve the existing assembly and gene annotations.

In summary, this work provides a set of biological resources, RNA-PET libraries providing better demarcation of genes, a list of alternative transcriptional start sites, promoter regions together with two large insert DNA-PET libraries (9 kb and 17 bp) helping re-assemble the published genome and providing the location of structural polymorphisms. The workflow presented demonstrates how these resources can be exploited by the scientific community for functional, structural and evolutionary analysis of genomes in non-conventional models.

Materials and Methods

Animals and treatments

Adult *X. tropicalis* frogs were obtained from NASCO (Fort Atkinson, WI) and maintained at 24°C in aquatic housing system (MPAquarien, Rockenhausen, Germany). Mating was induced by injection of 200U of human chorionic gonadotropine for females and 100U for males (Chorulon; Intervet, Beaucouze, France). Tadpoles were raised at 26°C. Tadpoles were staged according to the normal table of *Xenopus laevis* (Daudin) of Nieuwkoop and Faber (NF) (1967). For THs treatment, tadpoles at stage NF54 were exposed 24 h, in 5 liters with 10 nM T_3 (ref: T2752; Sigma, St. Quentin Fallavier, France). Tadpoles were killed by decapitation after anesthesia (ref: E10505; 0.01% MS222, Sigma) before hindlimb and tail fin dissection. Animal care was in accordance with institutional and national guidelines. Ethical approval for animal experimentation has been issued for this research project (ref: 68008, delivered by the Cuvier Ethic Committee).

RNA isolation

For each physiological condition, tissues were isolated from groups of 10 tadpoles, collected, flash frozen, and stored at -80°C. Tissue lysis was performed in 500 µl of RNAble (ref: Gex-ex-T00 – 0U; Eurobio, Les Ulis, France) with one bead (INOX AISI 304 grade 100 AFBMA) using Tissue Lyser II apparatus (Qiagen, Courtaboeuf, France) for 1 min at 30 Hz. The lysed tissues were mixed with chloroform and incubated on ice for 5 min before centrifugation (12,000 x g, 15 min, 4°C). The supernatant was subjected to RNA purification with RNeasy MinElute Cleanup kit according to manufacturer (ref: 74204, Qiagen). For Adult tissues, 2 to 50 frogs were required (based on the mass of tissue removed). Tissues were isolated, flash frozen, and stored at -80°C. Tissue lysis was performed in 5 ml of RNAble with an ultra turrax. The lysed tissues were mixed with chloroform and incubated on ice for 5 min before centrifugation (12,000 x g, 15 min, 4°C). The supernatant was subjected to RNA purification with RNeasy MidiElute Cleanup kit according to manufacturer (ref: 75142, Qiagen). RNA concentration was measured by optic density. RNA quality was estimated by microcapillary electrophoresis using Qiaxcel (Qiagen).

RT-qPCR analysis

Potential contamination by genomic DNA was removed using DNAse treatment as described by the provider (ref: AM1907; Turbo DNA free; Ambion, LifeTechnologies, Courtaboeuf, France). Reverse transcription (RT) was done as previously described (Bilesimo et al., 2011) using Superscript III reverse transcriptase (ref: 18080044; LifeTechnologies, Fisher Scientific, Illkirch, France). RT products were analyzed by real-time qPCR performed on an ABI 7300 (Applied Biosystems). Primers were designed using Primer express (Applied Biosystems). The list of used primers is given in Supplemental Table 1. Prism 7300 system software (Applied Biosystems) was used to analyze the results. Data are presented in means of Log(2ΔΔCT) and SEM (CT, cycle time). The raw data of three biological replicates are first normalized on the endogenous control gene rpl8 (ΔCT: mean CT rpl8 minus mean CT gene of interest) followed by Log transformation, so the variances between groups succeed the t-test. For statistical analysis, the Log of the normalized data was subjected to a one-sample two-tailed t test (〈= 5%).

Processing of RNA-Seq data

Conventional RNA-Seq was carried out on SOLiD plateform following the manufacturer instructions and reads were mapped using Bioscope pipeline. Paired-end RNA-Seq was carried out by SERVICEXS (www.servicexs.com, Leiden, Netherlands) on the Illumina HiSeq plateform. Reads were processed with Tophat (v2.0.6) and Cufflinks (v2.0.2) run with default parameters, except for the Tophat r parameter set to 100.

Construction of RNA-PET libraries

A 27-bp DNA tag sequence from each end of the full length cDNA was then extracted after a type III restriction enzyme (*Eco*P15I) digestion. The resulting Paired End diTag fragment (27bp-linker-27bp) was ligated to sequencing adaptors at both ends, amplified by PCR, and purified as templates for paired end (PE) sequencing with Illumina system. Construction of the RNA-PET genomic libraries (supplementary Figure 15) was using an improved and modified protocol version to the previous “GIS-PET”, which had been published and described in Ng et al. (2005*)*. Briefly, full length polyA mRNA were purified from high quality total RNAs with MACs polyT columns. Approximately one to two micrograms of purified mRNA were used for reverse transcription into full length cDNA, which were then biotinylated at the ends. Double stranded cDNAs were then circularized after ligation to specific DNA linker sequences. The 5’ and 3’ ends (27-27 bp) tags were then extracted through EcoP15I digestion and resulting PET templates [5’-27bp--linker—27bp-3’] which were purified from mixture by binding to streptavidin magnetic beads. Illumina sequencing adaptors were then ligated at both ends and PCR-amplified and sequenced by paired end reads at 2 x 36 bp long. Reads quality was assessed with standard quality controls. The modified version of the RNA-PET construction has two major changes: 1) it generates longer (27/27bp, in contrast to the previous 18/20bp) paired end tags from 5’ and 3’ ends of transcripts, by using a type III restriction enzyme EcoP15I, 2) eliminated bacterial-cloning procedures.

Processing of RNA-PET data

For each library, paired reads were independently mapped using bowtie version 0.12.7 (Langmead et al, 2009), with stringent parameters (l=22, n=0, m=1). After merging, mate pairs were subject to a stringent filter, so that only those for which both end map at a unique genomic location are retained. The orientation of transcript was resolved by looking for the three prime end-specific signature AACTGCTG in both mate pair with a modified version of the Smith and Waterman algorithm (Smith and Waterman, 1981). Only pairs with a 3' end signature at one end were kept for further processing. Clustering was carried out by using a greedy algorithm aggregating together PETs sharing overlapping 5' end tags. Models unsupported by RNA-Seq reads (<5 reads per kb) were discarded. The final models were further size filtered to reduce background noise originating from small artifactual models (see main text). All these were carried out with custom python scripts, cython and a C library.

Gene re-annotation

The annotation of the gene set was improved by integrating paired end RNA-Seq to Ensembl models, with the MAKER2 and EVM pipelines, following guidelines available from the respective web sites (http://gmod.org/wiki/MAKER, http://evidencemodeler.sourceforge.net/).

Antibodies

The TR antibody has been previously described for use in ChIP assay (Bilesimo et al. 2011). RNA PolII antibody (CTD4H8) was from Epigentec (A-2032; Euromedex, Souffelweyersheim, France).

ChIP and qPCR analysis

Up to 30 to 50 mg dissected tissues from seven euthanized tadpoles were used to isolate chromatin. ChIP was done as previously described with slight modifications (Bilesimo et al. 2011). ChIP products were analyzed by real-time qPCR as previously described (Bilesimo et al. 2011). The results were expressed as percent of input and presented as means SEM of at least three independent experiments.

Statistical analyses were performed with a paired two-tailed t test (α= 5%). The list of used primers is given in Supplemental Table 1. For ChIA-PET analysis, chromatin was isolated from a batch of 8 samples and subject to ChIP and ChIA-PET procedure (Fullwood et al., 2009). For the RNAPol II ChIP-Seq, ChIP was carried out on chromatin extract prepared from a batch of 20 samples. The ChIP product was purified by phenol chloroform extraction with Phase lock gel and ethanol precipitation. Sequencing of the RNA Pol II ChIP product was outsourced SERVICEXS (Leiden, Netherlands).

Preparation of genomic DNA

Genomic DNA was prepared from liver and kidney of seven male and female adults. Tissues were dissected and directly disorganized with Qiagen Tissue Lyser (30 Hz, 20 sec.) and the lysate was cleared with Qiagen QIAshredder columns (ref: 79654). Total DNA was extracted with DNA AllPrep DNA/RNA Mini columns (ref: 80204). All procedures were carried out according to the manufacturer's instructions.

Generation of DNA-PET libraries

To construct large insert DNA-PET libraries (Supplementary Figure 1), whole genome DNA was randomly hydrosheared, blunt end repaired, biotinylated and ligated to a specific DNA linker sequence at both ends. The linker-modified genomic DNA fragments were then gel-purified to obtain the size range (e.g., 10 kb, or 15 bp) of interest, followed by circularization with linker ligation. The paired end tags (PET at 27/27bp) were then excised through a type III restriction enzyme EcoP15I digestion. Extracted PETs were purified by binding to streptavidin magnetic beads, followed by ligation with SOLiD-specific adaptors at both ends. The complete sequencing template structure of the DNA-PET is amplified by PCR and processed for SOLiD mate-pair sequencing. Reads quality was assessed with standard quality control procedures.

DNA-PET data processing

Data cleaning and mapping: The two DNA-PET (mate-pair) libraries were mapped using the SOLiD Corona-Lite pipeline, allowing 2 mismatches for each 25 bp tag. The Corona-Lite mate-rescue step was carried out allowing up to maximum 4 mismatches in a read. For each library the mapping pipeline produces two different sets of mapping results. The ‘.map’ file contains all unique 5’ and 3’ mappings; the ‘.mates’ file contains possible unique 5’ – 3’ pairings. If a read cannot be paired concordantly and has a (discordant) unique 5’ and 3’ mapping, the discordant pair is listed in the ‘.mates’ file instead.

Scaffold splitting: We determined concordant PET range of IXT010 to be within 5997-11961 bp and that of IXT011 to be within 9318-22026 bp. We analyzed the concordant PET fragment coverage across the reference genome to determine possible mis-assemblies. At any instance where the concordant PET fragment coverage drops to 0 (a.k.a. assembly breakpoint), the scaffold was split into two different scaffolds. This resulted in 19,847 scaffolds. Furthermore, we split the resulting scaffolds at continuous blocks of 10,000 or more ‘N’s, provided there are no PET mappings spanning over these regions (21909 scaffolds).

Identifying repeat regions: Since the presence of repeated sequences can interfere with the scaffolding process, regions, which could have more multiple copies across the genome were identified (but assembled into a single region in the assembly). The assembly sequence was binned in 1000 bp windows, the number of unique 5’ and 3’ mappings in each window was counted. Any windows with more than 500 tags were treated as 'repeated regions' and were discarded.

Building graph: This step consists of bridging together scaffolds connected by sets of discordant PETs. Keep track that the resolution of this step is limited by the size of the DNA fragments used to build the libraries. A node in a graph represents each scaffold and DNA-PET discordant PETs are represented by an edge. When ≥ 3 discordant PETs connect the end of one scaffold to the end of another scaffold, they are linked with a weight equal to the total number of end-to-end discordant PETs they share. Each library (IXT010 and IXT011) was processed separately and they have their own distinct set of edges. For keeping track of the splitting process at assembly breakpoints, we introduced extra edges with weight 1 connecting the two scaffolds that were split at a blocks of ‘NNN’.

Scaffolding: An ordering of contigs is given a score based on how well it satisfies its edges. If the given ordering makes an edge “Concordant” (similar to PETs) the weight of edge is added to the total score of the ordering. If the edge is not concordant, relative to the given ordering, the weight of the edge is subtracted from the total score. The final contig ordering is found by trying to maximize the score for each scaffold. A heuristic algorithm carries this out. For further details of this algorithm, kindly refer to scaffolding section of PE-Assembler paper.

ChIA-PET analysis and processing of ChIA-PET data

The entire ChIA-PET analysis procedure consists of three main parts: ChIP sample prep, ChIA-PET library construction, and PET sequencing and mapping. Briefly, a typical ChIA-PET analysis starts with proper ages of cell culture; primary cells, or dissected tissues, which are cross-link fixed by proper chemical agent, and lysed for release of proper chromatin complex containing relationship between particular DNA region(s) and protein(s). Cross-linked chromatin complex (DNA/Protein) are fragmented by sonication to a given size range (~200-600bp), and enriched by chromatin immunoprecipitation (ChIP) method with a specific antibody of interest. In the enriched chromatin complexes, tethered DNA fragments (spatial located) are joined together by specifically designed DNA linkers through “proximity ligation”, and joined DNA fragment are then extracted for PET template, followed by high-throughput NGS sequencing. PET sequences are mapped to a reference genome, which can reveal genome-wide protein/DNA relationship of 1) binding sites (like CHIP-Seq data), and 2) long-range chromatin interaction relationships between two remote loci brought together by protein(s) at the interest.

Acknowledgments

We thank Dr N. Pollet for helpful discussions during the progression of this study. We also thank J.-P. Chaumeil and G. Benisti for animal care. We thank Zhenshui Zhang, Huay Mei Poh, Alexis KHNG Jiaying for DNA-PET, RNA-PET and Chia-PET construction, and GIS GTB sequencing group personnel Jo Ong, Yen Ling Lee, Dawn Choi Poh, See Ting for SOLiD and Illumina sequencing production).

This work was supported by the “Centre National de la Recherche Scientifique”, the “Muséum National d’Histoire Naturelle”, “CRESCENDO”, an Integrated Project funding from FP6 (contract n°LSHM-CT-2005-018652), “TRIGGER” (ANR-08-JCJC-0100-01) and “IDEAL”, a large Integrated Project funding from FP7 (contract n°259679). PB was supported by a “CIFRE” grant, Watchfrog and an ARC grant (“Association de Recherche contre le Cancer”). LS and YR were recipients of a Merlion Project (contract n°5.01.10).

Figure Legends

Figure 1. DNA-PET significantly improves genome assembly.

A, D. Fraction of each scaffold left unsequenced (expressed in per cent) as a function of size, in log scale. The initial genome assemble is plotted on the top, and the improved assembly of the bottom. The number of scaffolds chained together by dPETs ('chain length, in number of scaffolds) is indicated by blue, green and red colors. Purple circles denote the scaffolds corresponding to ~80% of the assembly. B. Raw DNA-PET data mapping statistics. C. Statistics of genome assembly improvement. E. Distribution of the number of scaffolds connected together with dPETs, showing that a large fraction of the total scaffolds are connected together into long chains. F. Example of re-scaffolding. A total of 15 scaffolds are linked together by dPETs. Tracks from top to bottom: scaffold name, assembly gap size, connectivity, number of dPETs per link for each DNA-PET library. The two scaffolds containing assembly gaps were split before re-scaffolding. Colored numbers indicate the number of independant dPETs supporting each connection, for each library.

Figure 2. RNA-PET efficiently captures transcripts ends.

A. Overlap between RNA-Seq reads and Ensembl and RNA-PET-based models. B. Demarcation of gene model boundaries by RNA-PET. The histogram shows the relative size of Ensembl gene models in bins of various sizes. C. Enrichment of RNA-Pol II around ensembl gene models and RNA-PET-based models. This shows that RNA-Pol II density fits well with RNA-PET based models, but not Ensembl models.

Figure 3. Examples of genome annotation improvements.

Track order: Ensembl models, RNA-PET based models, RNA-PET ditags, RNA-Seq reads density. A, B, C: sumo1, cadm2 and kiaa1958 loci. D: Un-annotated gene split over scaffold_1031 and scaffold_1460.

Figure 4. Benefit of genome re-annotation with RNA-PET for ChIA-PET analysis.

A. Large genomic view of the bcl6 locus. Track order: Ensembl genes, RNA-PET-based models, ChIA-PET TR binding density, interaction PETs, RNA Pol-II binding density, RNA-Seq reads density with (+T_3) and without (-T_3) treatment with thyroid hormones. B. Close up on TR binding sites. Track order: ensembl genes, RNA-PET-based genes, location of ChIP-qPCR probes, RNA-PET ditags, TR binding density, RNA Pol-II binding density. C: ChIP-qPCR validation of TR binding at locations shown in B. Ab: Antibody, T_3: 3',5,3' triiodothyronine treatment. D: Induction of trpg1, lpp and bcl6 genes transcription, assayed by RT-qPCR. E: Three-dimensional model of the locus topology.

Figure 5. Benefit of genome re-annotation with RNA-PET for ChIA-PET analysis.

A. Genomic view of an unannotated gene. Track order: Ensembl genes, RNA-PET-based models, ChIA-PET TR binding density, RNA Pol-II binding density, RNA-Seq reads density with (+T_3) and without (-T_3) THs treatment. B. Close up of TR binding sites. Track order: Ensembl genes, RNA-PET-based genes, location of ChIP-qPCR probes, RNA-PET PETs, TR binding density, RNA Pol-II binding density. C: ChIP-qPCR validation of TR binding at locations shown in B. Ab: antibody, T_3: 3',5,3' triiodothyronine treatment. D: Transcriptional induction, assayed by RT-qPCR.

References

Adelman,K. and LisJT., 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet 13:720-731

Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and SnyderM. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:pp.

57-74

Bilesimo ,P., Jolivet,P., Alfama,G., Buisine,N., Le Mevel,S., Havis,E., Demeneix,B.A. and SachsL.M. 2011. Specific histone lysine 4 methylation patterns define tr-binding capacity and differentiate direct T3 responses. Mol Endocrinol 25:pp. 225-237

Bombarely ,A., Rosli,H.G., Vrebalov,J., Moffett,P., Mueller,L.A. and MartinG.B. 2012. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol Plant Microbe Interact 25:pp. 1523-1530

Buchholz ,D.R., Paul,B.D., Fu,L. and ShiY. 2006. Molecular and developmental analyses of thyroid hormone receptor function in *Xenopus laevis*, the african clawed frog. Gen Comp Endocrinol 145:pp. 1-19

Buisine ,N. and SachsL. 2009. Impact of genome assembly status on Chip-Seq and Chip-Pet data mapping. BMC Res Notes 2:p. 257

Earl ,D., Bradnam,K., St John,J., Darling,A., Lin,D., Fass,J., Yu,H.O.K., Buffalo,V., Zerbino,D.R., Diekhans,M.et al. 2011. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Res 21:pp. 2224-2241

Eckalbar ,W.L., Hutchins,E.D., Markov,G.J., Allen,A.N., Corneveaux,J.J., Lindblad-Toh,K., Di Palma,F., Alföldi,J., Huentelman,M.J. and KusumiK. 2013. Genome reannotation of the lizard *Anolis carolinensis* based on 14 adult and embryonic deep transcriptomes. BMC Genomics 14:p. 49

Fullwood ,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H.et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462:pp. 58-64

Fullwood ,M.J., Wei,C., Liu,E.T. and RuanY. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.

Genome Res 19:pp. 521-532

Furey ,T.S. 2012. Chip-Seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. Nat Rev Genet 13:pp. 840-852

Gibcus ,J.H. and DekkerJ. 2013. The hierarchy of the 3D genome. Mol Cell 49:pp. 773-782

Graveley ,B.R., Brooks,A.N., Carlson,J.W., Duff,M.O., Landolin,J.M., Yang,L., Artieri,C.G., van Baren,M.J., Boley,N., Booth,B.W.et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:pp. 473-479

Grimaldi ,A.G., Buisine,N., Bilesimo,P. and SachsL.M. 2013. High-throughput sequencing will metamorphose the analysis of thyroid hormone receptor function during amphibian development. Curr Top Dev Biol 103:pp. 277-303

Grimaldi ,A., Buisine,N., Miller,T., Shi,Y.-B. and SachsL.M. 2013. Mechanisms of thyroid hormone receptor action during development: lessons from amphibian studies. *Biochim Biophys Acta* 1830: 3882-3892

Haas ,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and WortmanJ.R. 2008. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol 9:p. R7

Hellsten ,U., Harland,R.M., Gilchrist,M.J., Hendrix,D., Jurka,J., Kapitonov,V., Ovcharenko,I., Putnam,N.H., Shu,S., Taher,L.et al. 2010. The genome of the western clawed frog *Xenopus tropicalis*. Science 328:pp. 633-636

Holt ,C. and YandellM. 2011. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:p. 491

Hoskins ,R.A., Carlson,J.W., Kennedy,C., Acevedo,D., Evans-Holm,M., Frise,E., Wan,K.H., Park,S., Mendez-Lago,M., Rossi,F.et al. 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. Science 316:pp. 1625-1628

Huang ,Y., Chen,W., Wang,X., Liu,H., Chen,Y., Guo,L., Luo,F., Sun,J., Mao,Q., Liang,P.et al. 2013. The carcinogenic liver fluke, *Clonorchis sinensis*: new assembly, reannotation and analysis of the genome and characterization of tissue transcriptomes. PLoS One 8:p. e54732

Laudet ,V. 2011. The origins and evolution of vertebrate metamorphosis. Curr Biol 21:p. R726-37

Langmead ,B., Trapnell,C., Pop,M. and SalzbergS.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:p. R25

Li ,R., Fan,W., Tian,G., Zhu,H., He,L., Cai,J., Huang,Q., Cai,Q., Li,B., Bai,Y.et al. 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463:pp. 311-317

Li ,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J.et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148:pp. 84-98

Lin ,M.F., Carlson,J.W., Crosby,M.A., Matthews,B.B., Yu,C., Park,S., Wan,K.H., Schroeder,A.J., Gramates,L.S., St Pierre,S.E.et al. 2007. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. Genome Res 17:pp. 1823-1836

Mardis ,E.R. 2011. A decade's perspective on DNA sequencing technology. Nature 470:pp. 198-203

Nieuwkoop,P.D. and FaberJ. 1967. Normal table of *Xenopus laevis* (Daudin), 2nd ed., North-Holland Pub. Co., Amsterdam.

Ng ,P., Wei,C., Sung,W., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A., Wong,C.H.et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2:pp. 105-111

Reese ,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and LewisS.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. Genome Res 10:pp. 483-501

Ruan ,Y., Ooi,H.S., Choo,S.W., Chiu,K.P., Zhao,X.D., Srinivasan,K.G., Yao,F., Choo,C.Y., Liu,J., Ariyaratne,P.et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using paired-end ditags (PETs). Genome Res 17:pp. 828-838

Ruan ,X. and RuanY. 2012. Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). Methods Mol Biol 809:pp. 535-562

Sanyal ,A., Lajoie,B.R., Jain,G. and DekkerJ. 2012. The long-range interaction landscape of gene promoters. Nature 489:pp. 109-113

Schmutz ,J., Wheeler,J., Grimwood,J., Dickson,M., Yang,J., Caoile,C., Bajorek,E., Black,S., Chan,Y.M., Denys,M.et al. 2004. Quality assessment of the human genome sequence. Nature 429:pp. 365-368

Shi, 1999. Amphibian metamorphosis: From morphology to molecular biology (1st ed). Wiley-Liss, New York

Smith ,T.F. and Waterman M.S. 1981. Identification of common molecular subsequences. J Mol Biol 147:pp. 195-197

Tata ,J.R. 1993. Gene expression during metamorphosis: an ideal model for post-embryonic development. Bioessays 15:pp. 239-248

Tata,J.R. 2006. Amphibian metamorphosis as a model for the developmental actions of thyroid hormone. *Mol Cell Endocrinol* 246: 10-20

The modENCODE Consortium. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330:pp. 1787-1797

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Trapnell ,C., Pachter,L. and SalzberdS.L. 2009. Tophat: discovering smplice junctions with rna-seq. *Bioinformatics* 25: 1105-1111

Trapnell ,C., Williams,B.A., Pertea,G., Motazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and PachterL. 2010. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28: 511-515

Varshney ,R.K., Song,C., Saxena,R.K., Azam,S., Yu,S., Sharpe,A.G., Cannon,S., Baek,J., Rosen,B.D., Tar'an,B.et al. 2013. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:pp. 240-246

Warren ,W.C., Hillier,L.W., Marshall Graves,J.A., Birney,E., Ponting,C.P., Grützner,F., Belov,K., Miller,W., Clarke,L., Chinwalla,A.T.et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 453: 175-183

Wells ,D.E., Gutierrez,L., Xu,Z., Krylov,V., Macha,J., Blankenburg,K.P., Hitchens,M., Bellot,L.J., Spivey,M., Stemple,D.L.et al. 2011. A genetic map of *Xenopus tropicalis*. Dev Biol 354:pp. 1-8

Xu ,Q., Chen,L., Ruan,X., Chen,D., Zhu,A., Chen,C., Bertrand,D., Jiao,W., Hao,B., Lyon,M.P. et al. 2013. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet 45:pp. 59-66

Yandell ,M. and Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:pp. 329-342

Yao ,F., Ariyaratne,P.N., Hillmer,A.M., Lee,W.H., Li,G., Teo,A.S.M., Woo,X.Y., Zhang,Z., Chen,J.P., Poh,W.T. et al. 2012. Long span DNA paired-end-tag (DNA-PET) sequencing strategy for the interrogation of genomic structural mutations and fusion-point-guided reconstruction of amplicons. PLoS One 7:p. e46152

Zhou ,Q., Li,T. and Price D.H. 2012. RNA polymerase II elongation control. Annu Rev Biochem 81:pp. 119-143

Zhuo ,Y., Liu,L., Wang,Q., Liu,X., Ren,B., Liu,M., Ni,P., Cheng,Y. and Zhang L. 2012. Revised genome sequence of *Burkholderia thailandensis* msmb43 with improved annotation. J Bacteriol 194:pp. 4749-4750