Eléments transposables et génomes eucaryotes

Hadi Quesneville

Université Pierre et Marie Curie

Paris VI

Dossier d'Habilitation à Diriger des Recherches

2004

Dossier de recherche

Présenté par

Hadi Quesneville

REMERCIEMENTS

Je tiens à remercier Dominique Anxolabéhère pour ses encouragements et la confiance qu'il m'a témoignés tout au long de ces années.

Merci à Danielle Nouaud pour son soutien et son aide durant toutes ces années.

Un grand merci aux rapporteurs qui ont accepté de lire ce manuscrit malgré leurs emploi du temps surchargés.

Mille mercis aux étudiants que j'ai pu encadré pour ce qu'ils m'ont apporté humainement et scientifiquement. Le travaille présenté ici leur doit beaucoup.

Enfin merci aux membres du laboratoire Dynamique du Génome et Evolution qui m'ont accompagnés durant toutes ces années.

TABLE DES MATIERES

1. RESUME

Le séquençage systématique des génomes révèle que les éléments transposables (ET) représentent à eux seuls une proportion importante de leur séquence (45% pour le génome humain). Longtemps perçus comme des parasites intra-génomiques, de récents résultats suggèrent qu'ils sont à l'origine de nombreuses fonctions biologiques essentielles (Kidwell et Lisch, 2001). Ils sont aujourd'hui perçus comme un des facteurs clef de la dynamique des génomes et de leur évolution, mais aussi de leur fonctionnement et de leur structure. L'analyse de l'organisation des séquences mobiles, ou ayant été mobiles, au sein d'un génome est devenue aujourd'hui possible. Elle se révèlera très fructueuse pour comprendre la dynamique évolutive des génomes.

L'annotation correcte d'un génome et de ses ET est un préalable indispensable à cette étude. L'approche classiquement utilisée consiste à comparer par alignement les séquences génomiques avec les séquences connues d'ET. Basés sur cette approche, nous avons développé des outils d'annotations particulièrement bien adaptés à la détection des ET. La comparaison de nos annotations du génome de *D. melanogaster* (release 3) à celles réalisées par BDGP (Berkeley Drosophila Genome Project) disponibles sur Flybase, montre que nos outils sont beaucoup plus sensibles.

Cependant, cette approche ne permet de détecter que des éléments connus, ou des séquences proches de ceux-ci. Nous avons donc développé des méthodes permettant de chercher de nouveaux ET dans les génomes. La première approche recherche les séquences codantes similaires à celles de différents éléments connus. Généralement la séquence codante des éléments de références n'est que partiellement ou pas connue. Notre stratégie a donc été de comparer les traductions dans les 6 phases des séquences génomiques avec celles des ET actuellement répertoriés dans la banque de données « REPBASE UPDATE» dédiée aux éléments répétés des génomes eucaryotes. Nous pouvons alors comparer des fragments de séquences codantes, et ainsi rechercher des similitudes au niveau protéique, même avec des séquences partielles. Le nombre de séquences d'ET identifiables est alors considérablement augmenté : notre méthode de détection est très sensible. Ainsi nous avons pu mettre en évidence plus de 300 ET non encore détectés chez la Drosophile et 400 pour l'Anophèle (Quesneville et al. 2003).

Cependant, cette dernière approche reste encore très dépendante des séquences d'ET connus. Nous avons alors développé également une approche basée sur les biais de composition nucléotidique (Andrieu O, Fiston A, Anxolabéhère D, Quesneville H, 2004). A l'aide de modèles de Markov cachés (HMM), nous avons étudié ces biais sur trois organismes: *D. melanogaster*, C. *elegans* et *A. thaliana*. Avec les HMM, nous montrons que les ET complets de classe I et II ont une composition qui permet de les distinguer des gènes et qu'il existe une différence entre ces classes d'ET. Nos tests chez *D. melanogaster*, *A. thaliana*, et *C. elegans* donnent des sensibilités de détection approchant les 70 % et des spécificités approchant les 100 %, indiquant que l'utilisation des HMM représente un moyen performant de détection et d'annotation des séquences d'ET.

Les blocs dupliqués dont la taille est comprise entre 1 et 200 kb sont appelés duplications segmentales (DS). Nous avons réalisé une étude *in silico,* dans le génome de *Drosophila melanogaster,* de la structure des DS et cherché l'implication des ET dans celles-ci. Nous avons pu mettre en évidence la présence fréquente de fragments d'ET au niveau d'une seule de leurs extrémités. Nous avons proposé un modèle expliquant la structure et la dynamique de ces DS. Ce modèle propose que l'homologie de séquence des fragments d'ET, serait en partie responsable de la formation des DS grâce à un mécanisme de conversion génique entre séquences homologues non-allèliques (SDSA), mécanisme employé lors de la réparation des cassures doubles brins de l'ADN. Nous avons identifié sur les séquences dupliquées les traces caractéristiques laissées par le modèle SDSA.

En parallèle, nous avons développé un environnement de simulation orienté-objets, appelé GENOOM, qui propose un nouveau cadre conceptuel pour l'étude de l'évolution des séquences d'éléments transposables. Inspiré d'une approche « vie artificielle », nous considérons les copies d'un élément transposable sous la forme de séquences binaires insérées dans le génome d'individus constitués en populations. Les chromosomes des individus sont soumis aux opérateurs génétiques classiques tels que mutations et recombinaisons, mais aussi aux opérateurs propres au système modélisé tels que la transposition et la réparation. La survie des individus étant affectée par la mobilité des éléments, le système évolue vers un équilibre. Ce modèle est implémenté dans l'environnement de simulation que nous avons construit. Nous pouvons ainsi explorer les propriétés des séquences sélectionnées, mais aussi leurs distributions au sein des génomes

et des populations. Nos résultats montrent qu'un élément transposable peut émerger à partir d'un gène unique et immobile capable de réaliser des coupures de l'ADN. Puis ce nouvel élément lorsqu'il se multiplie par transposition, exerce une pression de sélection qui favorise l'apparition de nouvelles copies régulatrices limitant la mobilité de celui-ci. Les interactions entre l'hôte, les copies actives et les copies régulatrices générées au cours de la simulation provoquent une évolution concertée des différentes régions de l'élément transposable. Nous avons étendu les capacités de simulation de notre simulateur dédié aux éléments transposable. Il permet de simuler des populations spatialement réparties, à effectifs variables, avec un mélange de classes d'éléments génétiques (microsatellites, marqueurs RFLP, gènes, QTL, …), une stérilité liée à certains phénotypes, des croisements entre apparentés, et des migrations. Cet outil est actuellement utilisé à l'unité INSERM 535 pour tester diverses méthodes d'analyses des populations humaines (Bourgain et al. 2000, 2001a, et b, 2002).

Après le séquençage de *D. melanogaster* et *D. pseudoobscura,* 2 nouveaux génomes de Drosophiles sont séquencés : *D. yakuba* et *D. simulans*. Les séquences génomiques de ces 4 Drosophilidés ouvrent de nouvelles perspectives pour l'étude des remaniements chromosomiques. En effet, lorsqu'on ne regarde qu'un seul génome, on ne peut détecter que les évènements de duplications. Tous les évènements de délétions, inversions, et translocations passent inaperçus. Le séquençage de *D. yakuba* et de *D. simulans*, ouvre la voie à ce type d'étude. On dispose de génomes suffisamment proches phylogénétiquement pour conserver de grandes régions synthéniques. Celles-ci nous permettrons de détecter tous les types de remaniements chromosomiques et d'approfondir l'étude de notre modèle de duplication. Collaborant pour l'annotation des ET de ces génomes, nous seront tout particulièrement bien placés pour aborder l'impact des ET sur ces remaniements. Nous chercherons à déterminer le rôle des ET dans ces remaniements : leur importance et les mécanismes.

Enfin nous voulons étudier l'impact des produits des ET sur le transcriptome. Pour cela nous envisageons une étude différentielle de transcriptomes de lignées avec ou sans éléments *P*. L'étude sera réalisée par une approche par puce à ADN, mais aussi en recherchant, par détection *in silico*, des gènes présentant dans leur environnement génomique des séquences ressemblant aux sites de fixation de la transposase de *P*.

2. EXTENDED ABSTRACT

My work has essentially focused on studying at different levels eukaryotic transposable elements from molecular to population and trans-specific levels. I analyze interactions between transposable elements and the host genome and their consequences on the evolution of populations and species.

Transposable elements (TEs) are mobile DNA sequences that can be found repeated and dispersed in a genome. All the copies derived by transposition from a TE sequence belong to what is called a TE family. They can be classified according to the mechanism by which they move from one genomic site to another. Class I TEs correspond to elements that use an RNA intermediate in their transposition; they are also called retrotransposons. Class II TEs, known as DNA-transposons or DNA TEs, use DNA. Within a class II TE family, transposases provided by some intact copies can mobilize defective elements.

TEs have been found in nearly all genomes where they have been sought. They seem to be ubiquitous, and represent a quantitatively important component of genomes (44.4% of the human genome, International Human Genome Sequencing Consortium, 2000). Since the vast majority of the mutations caused by transposons are deleterious, their presence in the genome does not provide any apparent advantage for the survival of the host organism. In spite of their negative effect, transposons can spread and persist in the gene pool of a species over long periods of time.

2.1. TEs ANNOTATION

Transposable elements (TEs) represent a quantitatively important component of genomes. In *D. melanogaster* they represent ~20% of the whole genome (~5% of the euchromatin and ~50% of the heterochromatin). The procedures generally used to detect TEs in nucleic acid sequences rely on sequence similarity search with previously characterized elements (RepeatMasker, Censor, …).

TE copies in a genomic sequence can carry deletions over most of their sequence. They are numerous and fragmented. Sometimes copies are also nested in other copies,

making them difficult to annotate accurately. We have developed strategies and tools specifically designed to annotate TEs in genomes, taking into account their fragmented nature and their possible nested structures. We wrote 3 softwares: BLASTER, MATCHER, and GROUPER.

BLASTER is a C++ program that can be used to annotate TEs in genomic sequences. It can compare two sets of sequences: a query databank against a subject databank. For each sequence in the query databank, BLASTER launches one of the BLAST programs (blastn, tblastn, blastx, tblastx, blastp, megablast) (Altschul et al., 1990, 1997) to search the subject databank. The program is able to launch each BLAST search in parallel on a computer cluster using the PVM library. BLASTER is not limited by the length of sequences. It cuts long sequences before launching BLAST and re-assembles the results afterwards. Hence, it can work on whole genomes, in particular to compare a genome with itself to detect repeats. The results of BLASTER can then be treated by the MATCHER and GROUPER programs.

MATCHER is another C++ program that we developed to treat BLASTER results and to map the matches (HSPs) of the subject sequences on the queries. Cross hits are filtered as overlapping hits. Here, overlapping hits means also included. So when two HSPs overlap on the genomic sequence, the one with the best alignment score is kept, the other is truncated such as non-overlapping region remains on the HSP. As a result of this procedure an HSPs is totally removed only if it is included in a longer one with a better score. By this way, nested elements are kept. Long insertions (or deletions) in one of two homologous sequences result in two HSPs, instead of one with a long gap. To overcome this drawback, the remaining HSP are chained by dynamic programming as in the FASTA algorithm. A score is calculated by summing HSP scores and substracting a gap penalty and a mismatch penalty.

GROUPER is another C++ program that we developed to treat the BLASTER results. It uses HSPs (or chained HSPs) to gather similar sequences into groups by simple link clustering. An alignment belongs to a group if one of the two aligned sequences already belongs to this group over 95% of its length. Groups that share sequence locations are regrouped into what we have called a cluster. As a result of these procedures, each

group contains sequences that are homogeneous in length. A given region may belong to several groups, but all of these groups belong to the same cluster

Using BLASTER and MATCHER we have re-annotated the *D.melanogaster* release3 and compared its results to the official annotation from the annotations given by the Berkeley Drosophila Genome Project (BDGP). To compare our performances, we have also compared the official annotation with those obtained by the RepeatMasker software (http://repeatmasker.genome.washington.edu), classically used for TE annotations.

The results show that our tools are more accurate than RepeatMasker: the boundaries are better identified, and the fragmented and nested structures, better recovered. Moreover we show that many small copies are missing from the official annotation

When studying sequences from an organism which TEs are not well described, usual techniques for detection of TEs in nucleic sequences are likely to miss many "unknown" elements, as they rely on sequence similarity with already characterized elements. Indeed, by comparing the six-frame translations of the nucleic acid sequences of known TEs with the *D. melanogaster* genomic sequence, we have detected more than 300 potentially new TE families (Quesneville *et al.* 2003). We then propose a strategy that could hopefully detect "unknown" elements, *i.e.* that operates without *a priori* knowledge over the TEs that are to be found in the sequence.

Previous work has shown that TEs tend to have a different nucleotide composition than the host genes, either considering codon usage bias or dinucleotide frequencies. We show how these compositional differences can be used as a tool for detection and analysis of novel TE sequences (Andrieu *et al.* 2004). We compared the composition of TE sequences and host gene sequences using probabilistic models of nucleotide sequences. We used hidden Markov models (HMM), which take into account the base composition of the sequences and the heterogeneity between coding and non-coding parts of sequences. We analysed three sets of sequences containing class I TEs, class II TEs and genes respectively in three species; *Drosophila melanogaster*, *Cænorhabditis elegans* and *Arabidopsis thaliana*. Each of these sets had a distinct, homogeneous composition, enabling us to distinguish between the two classes of TE and the genes. However the particular base composition of the TEs differed in the three species studied. This approach can be used to detect and annotate TEs in genomic sequences and complements the current homology-

based TE detection methods. Furthermore, the HMM method is able to identify the parts of a sequence in which the nucleotide composition resembles that of a coding region of a TE. This is useful for the detailed annotation of TE sequences, which may contain an ancient, highly diverged coding region that is no longer fully functional.

Using this approach on genomic sequences, we identify new TE families even in species where the TEs are well studied. In order to identify and build consensus sequences for new TEs in a genomic sequence, we tested a two steps strategy called hereafter BLASTER-HMM:

1. We search for TE sequences by both comparing the six-frame translations of the nucleic acid sequences of known TEs with these of the genomic sequence of interest. The detected regions correspond generally to the conserved coding parts of the TEs. In order to find their non-coding sequences, the flanking sequences of these hits are compared all together by alignment. Flanking sequences that are similar (90% at nucleic acid level) between several hits are supposed to correspond to the rest of the TE sequence. Finally TE consensus sequences are built after multiple alignments of these similar regions.
2. We search in the genome, regions that have a base composition close to that of TEs using Hidden Markov Models (HMM). The HMM is trained on the consensus sequences obtained in the first phase. The resulting regions are then aligned in order to give new consensus as described above.

These two steps use tools that we have developed as follows:

1. Search for TE sequences hits by:
    a. both comparing the six-frame translations of the nucleic acid sequences of known TEs with these of the genomic sequence of interest, using BLASTER with *tblastx*.
    b. or using HMM to detect nucleotid composition close to that of TEs
2. Extend hits taking 5000bp at each side *(only for tblastx matches)*
3. Compare all together by BLASTER with *blastn* the (extended) sequences
4. Cluster matches that are similar by 90% of identity using GROUPER

5. TE consensus sequences are built as follows:

    c. Obtaining the multiple alignments of the sequences in the groups given by GROUPER

    d. From the multiple alignments, determining the consensus sequences taking the bases with frequency by column > 0.5 (not counting gaps) otherwise put 'N'.

    e. Taking for each cluster, the consensus sequences of its groups and align them.

    f. Determining the consensus sequence from the multiple alignment of the group consensus as before.

In order to evaluate this strategy, we have realised a blind test with *D. melanogaster*. We have analysed the *D. melanogaster* genome as if we don't know its TEs. Then, we have compared: (i) the obtained annotations with a "reference annotation" obtained by BLASTER and MATCHER using the known TE sequences available at BDGP web site, and (ii) the canonical TE sequences after building their consensus.

We have also compared our performances with those from two published methods RECON (Bao and Eddy, 2002) and LTR_STRUC (McCarthy and McDonalds, 2003):

- RECON works by clustering repeats obtained by an all-by-all sequence comparisons (*e.g.* with blastn) and redefined the clusters by the aggregation of endpoints in the multiple alignment of the identified regions
- LTR_STRUC searches for LTR retroelements. It seeks for its LTR by searching by alignment a repeated segment supposed to correspond to the LTR.

Our results show for each tested methods ~85% of the nucleotides from the "reference annotation" overlap the results given by either BLASTER/HMM, RECON or LTR_STRUC (only LTR TEs are considered for LTR_STRUC comparison). BLASTER-HMM detects many more potentially new TEs (identified as TE but not corresponding to known ones) than the other methods. RECON allows a better recovery of the complete sequence of the canonical TEs after consensus building. But it also gives many more redundant and partial consensus. The 3 methods are highly complementary as some TEs are specific of each one: 27 complete TEs are recovered by only one method and 98% of

the nucleotides from the “reference annotation” overlap the TE annotations obtained with the 3 methods combined!!

Using a combination of the 3 methods we obtained in *D. melanogaster* 363 consensus corresponding to potentially new TEs

2.2. TEs AND SEGMENTAL DUPLICATIONS

Long sequence duplications over more than several kilobases are a common feature of eukaryote genomes (Bailey *et al.* 2001, Horvath *et al.* 2000a and b, Samonte *et al* 2002). They are called segmental duplications (DS). The mechanisms that generate these long duplications remain unknown.

Due to their ability to cause rearrangements, TEs appear to potentially being able to provoke these duplications, either acting as short dispersed homologous sequences inducing unequal exchanges, or directly when they transpose (Gray 2000)

We have searched segmental duplications in the *D. melanogaster* genome to investigate if TE could be involved in their formation and what molecular mechanism could be involved. The DS have been search using BLASTER in an all-by-all genome comparison and GROUPER. The repeats detected have then been sorted out according to our BLASTER TE annotation, to remove repeats that correspond to TE copies. To be more stringent, we have selected the repeats covered by less than 95% by TEs: our DS have at least 5% of their sequence that differs from a TE.

We have detected 101 DS with up to 10 copies.

- As in the human genome, they are located preferentially in the pericentromeric heterochromatin, but not in the subtelomeric regions. The fourth chromosome arm, known to be mainly heterochromatic, has a DS density 3-fold higher than other chromosomes. These results suggest a DS preference for heterochromatic regions.
- 88% of the duplications are intra-chromosomique
- More than 75% of the DS have 2 copies.
- More than 3/4 of the copies have 100% of sequence identity

- 50% of the copies are distant by less than 7 kb
- The longest DS is 50 kb long

Then we have searched if TEs could be involved in their formation. Fifty eight percent of the copies have at least one TE at a boundary, and in mean, DS are covered by TE at 38%. We have proposed a model of formation that may explain this feature. It proposes that DS are formed by a double strand break repair mechanism using a TE copy as template for initiating the conversion.

We have tested this model by comparing the genomic distribution of TE to that of DS. It appears that the density per chromosome is significantly different. A re-sampling procedure shows that the high TE coverage percentage of DS and the high TE copies occurrence at their boundaries are not different from those obtained at random, when re-sampling in the same genomic regions. This indicates that TEs are not involved in DS formation, but have an affinity for the same regions.

We have also searched into the DS sequences, traces of the "gap repair" mechanism, a model proposed by Nassif *et al.* (1994). This model, first proposed to explain deletions induced by *P* transposons, could leave on the neo-synthesised strand small tandem repeats (up to 10 nucleotides). More than 70% of the DS have these traces, suggesting that this repair mechanism may be involved in the DS formation.

2.3. DYNAMICS OF TRANSPOSABLE ELEMENTS

Intense speculation on how these elements are maintained and spread by virtue of their transposition processes has produced a great deal of theoretical work on their evolutionary dynamics (Charlesworth & Langley, 1989; Badge & Brookfield, 1997; Brookfield & Badge, 1997; Quesneville & Anxolabéhère, 1997a, 1998, 2001).

The study of genetic interactions between TEs and their consequences on genome plasticity has allowed us to develop a theoretical approach of the evolutionary life history of TE families. Using a modelling approach, we have studied the *P* transposable element dynamics (Quesneville and Anxolabéhère 1997a, Quesneville and Anxolabéhère 1998). We have focused our studies on the formalization and the integration of knowledge to test if our current understanding of *P* transposable elements is consistent from the molecular

level up to the species point of view. That is to say: do experimentally-identified mechanisms explain *P* invasion at the population and meta-populations levels ? We have first integrated the various mechanisms of regulation and the transposition events already identified experimentally. These mechanisms are described and represented at a cellular level, in order to observe the emergent behaviour of the system at a species level, that is to say, the spread of the *P* transposable element in *Drosophila* natural populations. The model allows to obtain by simulation the different states observed in natural population. The simulations have also shown that the main actors of the *P* dynamics are: (i) the occurrence of migration (ii) the efficiency of DNA repair mechanisms in the host, and (iii) and the position effect of the *P* element insertion sites.

Recently, several cellular functions have been found to be closely related to TEs. For example, telomerase is a reverse transcriptase that is related to that of non-LTR retrotransposons. It is not clear whether retrotransposons gave rise to telomerase or *vice versa* (Nakamura & Cech, 1998; Pardue *et al.,* 1997). Another spectacular example is the origin of antigen-specific immunity. The V(D)J recombination system shares two main features with DNA transposons: a recombinase encoded by *RAG1* and *RAG2*, and a mobile DNA sequence flanked by recombinase-binding sites. This led to the hypothesis that the immune system comes from the insertion of a TE in the genome of an ancestor of a jawed vertebrate (Agrawal *et al.*, 1998; Hiom *et al.*, 1998). Numerous other examples can be found in Smit (1999).

Very little is known about the origin of TEs, but it is conceivable that it can be traced back to the hypothetical transition from RNA-based genomes to DNA-based ones (Jurka, 1998). There is no doubt that the genomic DNA we observe today has evolved with the close participation of TEs. Many of them have evolved into parasites, but they have probably all kept their properties of "genome builders". They appear today as crucial actors of genome evolution. Class II TE transposases have endonuclease properties: they achieve site specific DNA cleavages (Beall & Rio, 1997). Thus class II TEs can be thought of as endonuclease genes that recognize the extremities of their own nucleotide sequence. Consequently, these TEs could derive from this kind of genes.

The second point we have explored was whether TE family can emerge from gene with basic endonuclease properties (Quesneville *et al.,* 2001). To search for the ancestral

TE structure, we looked for a minimal organization common to all class II TEs. But, if the first transposase was an endonuclease, its mobile gene needs to generate copies that control their mobility in order to be maintained as a TE family. Thus, by the action of protein-DNA interactions and DNA repair, we expect this structure to develop a control mechanism for its mobility.

We have proposed a new conceptual framework inspired by recent approaches used to study “complex adaptive systems“. Based on a specifically designed genetic algorithm, the emergence of a class II TE family is studied as a self-organizing system.

Our results show that a TE could emerge from a single gene with basic endonuclease properties. Upon transpositions, this new element can produce mutated and deleted copies. Some of them could interact with other copies in order to reduce the invading capacity of the element. We observe the spontaneous formation of an organized molecular interaction network leading to a controlled mobility. The DNA double strand break repair process plays an important role in the invasion dynamics, and it appears to condition the emergence success of class II TEs. The distribution of the deletions in the sequences is not centred on the middle of the copies as could be expected given the repair process, but it is shifted to the "endonuclease domain". The deletion distribution also affects the rate of evolution of the sequence. An antagonism between two selective forces gives rise to heterogeneity along the TE sequences but also between the different copy types.

2.4. OBJECT-ORIENTED MODELING OF COMPLEX GENETIC SYSTEMS

The modelling we have used for such complex genetic systems requires a way of integrating a heterogeneous hierarchy of models (differential equations, stochastic processes, discrete events, markov chains, etc). This is because in such "multimodels", each model component at a given "level of abstraction" (e.g. a metabolic model at the individual level) may be sub-refined into other models, possibly of a different type, at the next lower level (e.g. enzymatic models at a molecular level). Multimodels require modellers to find ways to formalize their conceptual models and in order to integrate them; they would be able to navigate through successive levels of abstraction. The Object-Oriented Design offers a convenient modelling method integrating all the modelling steps, from the description of the conceptual model to its execution with a simulation program.

Because digital objects can be observed at different abstraction levels, the analyst may study the dynamics of a system at any chosen specific level.

We have designed such modelling tools to study TE dynamics (Quesneville and Anxolabéhère, 1997a, 1997b, 1998, 2001). Based on an object paradigm, they implement a virtual computer world where biological entities are digital objects. In this world, each individual of a population is represented. These tools allow studying complex genetic models by simulations performed according to a genetic map. They provide a framework of object classes where modellers can integrate any code. The geneticist implements the model he wants to simulate by writing the C++ code of its genetic model and plug-in it into the appropriate class provided by a C++ library. Then according to this model, a simulator will provide simulated populations from which various analyses could be performed.

We have adapted a tool, called GENOOM for GENetic Object-Oriented Modelling to model human populations according to the various models in epidemiology genetics. We have implemented different kind of genetic markers, parameters such as penetrance matrix (phenotype probabilities for each genotype), exposure to an environment, reproductive rate, inter-relative mating probabilities, and migration in a two-dimensional space. A sampler program comes with GENOOM. It allows the sampling of individuals in the populations produced by the simulation program. The sampler can randomly draw individuals from a same location or draw pedigrees with individuals sharing a particular phenotype. Procedures can, for example, sample nuclear families for sib-pairs analysis, pedigrees with first cousins, or larger families over several generations. The produced data files could be directly analyzed with different packages for statistical or genetic analysis such as LINKAGE (Terwilliger and Ott, 1994), GENEHUNTER (Kruglyak, *et al*, 1996), MAPMAKER/SIBS (Kruglyak and Lander, 1995). Pedigrees can also be simulated according to a pedigree file in LINKAGE format.

Simulation approaches play an important role in the development of new statistical methods in epidemiology genetics. Currently available programs include SIMLINK (Boehnke, 1986), SLINK (Weeks *et al,* 1990), SIMULATE (Terwilliger and Ott, 1994), and GASP (Wilson *et al,* 1996). They simulate the genotypes of families under a defined structure and linkage parameters. However, in searches of multifactorial disease susceptibility genes, the methods tend to take advantage of population features such as

founder effects, small population sizes, or consanguinity. Such type of analysis requires programs able to simulate populations with such features. Available programs such as POPSIM (Hampe, 1998) specifically dedicated to population studies, do not allow to simulate all these features. GENOOM allows performing these simulations. This package has been used in various population genetic analysis and genetic epidemiology studies, in particular to study founder population properties in the search for multifactorial disease susceptibility genes (Bourgain, *et al.*, 2000, 2001a and b, 2002).

2.5. PROJECTS

Our project concerns the impact of TEs on the genomes. Three complementary axis will be developed: (i) the detection and the identification of TEs in genomic sequences, (ii) a modelling of the genome evolutionary dynamics driven by TEs, and (iii) the impact of the TE product on the transcriptome.

2.5.1. TE annotation

We have set up several international collaborations to annotate TEs in genomes.

1. A first collaboration with Michael Ashburner and Casey Bergman (University of Cambridge) aims to annotate the *D. melanogaster* genomic sequence release 4. This work is in progress.
2. Two new Drosophila genomes have been sequenced: *D. yakuba* and *D. simulans*. We have planed with Michael Ashburner and Casey Bergman to annotate their TEs.
3. The *D. melanogaster* releases 1 to 4 only concern the euchromatic part of its genome. The sequencing of the heterochromatin is still in progress. With Michael Ashburner and Casey Bergman we have contacted Gary Karpen (Lawrence Berkeley National Laboratory in Berkeley) to annotate this genome compartment. Our tools being able to describe nested TE structures, they appear well suited to annotate TEs in these regions.
4. Another collaboration is started to annotate TEs in the *Anopheles gambiae* genome. This collaboration involves: Charles Roth (Institut Pasteur, France), Paul Brey (Institut Pasteur, France), Zhijian Tu (Virginia Polytechnic Institut

and State University, USA), Frank Collins (University of Notre Dame, USA), and the Ensembl team (EMBL-EBI et Sanger Institut, Royaume Uni).

For these projects we need to improve our TE annotation pipeline. Our pipeline will be developed as follows:

1. Installation and development of a set of software able to identify new TEs (*i.e.* BLASTER-HMM, RECON, LTR_STRUC, …), to detect TE structures in genomic sequences (*i.e.* BLASTER, RepeatMasker, Censor, …), to visualise and annotate predictions (*i.e.* Apollo, GBrowse), and to compare different sources of predictions to automatically promote annotations.
2. A database able to stock sequences, available annotations, and data produced by our software.
3. Software that schedule jobs, distribute them on a computer cluster, and control their execution.

2.5.2. Chromosomal rearrangements

It has been shown that TEs are able to induce chromosomal rearrangements such as deletions, duplications, inversions, or translocation (Gray 2000). We have shown that segmental duplications in *D. melanogaster* can be formed by the "gap repair" mechanism proposed by Nassif *et al.* (1994), using short non-allelic homologous regions to initiate the synthesis. The 4 sequenced Drosophila genomes (*D. melanogaster, D. pseudoobscura, D. yakuba*, and *D. simulans*), will allows us to explore more deeply this process. In particular, by looking at the synthenic regions of these duplications, it becomes possible to determine which copy corresponds to the neo-synthesized strand. This allows us to better test the model by looking at, specifically on the new strand, the tandem repeats predicted by the repair model. Moreover, the history of these duplications could be resolute, and then we could observe their evolution.

By studying synthenic regions between these genomes, it could be possible to observe the other rearrangement types such as deletions, translocations and inversions. The history of these events could be also determined and observed at different time scales. Being involved in the TE annotation of these genomes, we shall be in good position to determine the impact of TEs on these rearrangements.

Once a model validated from these data, we could improve GENOOM to be able to simulate these rearrangements, and study their dynamics at a population level.

2.5.3. TEs products impact on the transcriptome

We postulate that TEs products being able to bind the DNA, could affect the transcription of adjacent genes. Several observations on P-enhancer-traps, show that their expression is sensitive to the regulation state (P cytotype) of the P-elements. Such interactions could explain the evolutionary success of several P-element domestications (Nouaud *et al.* 2003). Our model to study these interactions will be the P-element. One important point with this TE, is the existence of *D. melanogaster* strains totally devoid of any P-element sequences that could be used to observe the transcriptome in the absence of P-elements.

The binding site of the P-transposase has been identified and a consensus of 10 pb has been proposed. Considering the numerous P-sequences available today, it is possible to refine it. This consensus can be then searched on the genomic sequence to identify genes that are close to a putative P-transposase binding site. The candidate genes can then be tested by quantitative RT-PCR, for an altered transcription when in presence of a P-element product.

Another approach is a microarray assay. By a differential analysis between the transcriptome of a line devoid of *P*-element and the same line with *P*-elements (such lines, obtained by transformation, already exist), it could be possible to observe genes up or down regulated. The candidate genes must be then tested as above for their altered transcription.

3. TRAVAUX RÉALISÉS

Mon travail s'est principalement orienté vers l'étude des éléments transposables eucaryotes, depuis le niveau moléculaire jusqu'au niveaux populationnels et trans-spécifiques. J'ai analysé les interactions entre les ET et le génome de leurs hôtes, ainsi que leurs conséquences sur l'évolution des populations et des espèces.

Depuis quelques années les séquences génomiques de nombreuses espèces modèles sont disponibles. Depuis peu, des séquences génomiques appartenant à des espèces proches de nos génomes modèles, ou encore plusieurs séquences génomiques d'une même espèce deviennent disponibles. Beaucoup de nos « vieilles » questions évolutives sont aujourd'hui abordées sous un angle global plus ambitieux. La génomique ouvre de nouvelles voies et de nouvelles façons de répondre à d'anciennes préoccupations.

Le séquençage systématique des génomes révèle aujourd'hui que les éléments transposables (ET) représentent à eux seuls une proportion importante de leur séquence (45% pour le génome humain). Les ET bénéficient ainsi d'un regain d'intérêt tout à fait remarquable. Longtemps perçus comme des parasites intra-génomiques, de récents résultats suggèrent qu'ils sont à l'origine de nombreuses fonctions biologiques essentielles comme le système immunitaire, les centromères, ou encore les télomérases (Kidwell et Lisch, 2001). Ils sont aujourd'hui perçus comme un des facteurs clef de la dynamique des génomes et de leur évolution, mais aussi de leur fonctionnement et de leur structure. L'analyse de l'organisation des séquences mobiles, ou ayant été mobiles, au sein d'un génome est devenue aujourd'hui possible. Elle se révèlera très fructueuse pour comprendre le fonctionnement et la dynamique évolutive des génomes.

Notre travail s'inscrit déjà dans cette dynamique, et nos projets s'orientent vers l'exploitation des nouveaux outils qui émergent autour de ces données génomiques.

3.1. ANNOTATION DES ELEMENTS TRANSPOSABLES DANS LES SEQUENCES GENOMIQUES

3.1.1. Méthodes

La détection des éléments transposables dans les séquences génomiques se place en amont de toutes ces études. En effet l'annotation correcte d'un génome et de ses ET est un préalable indispensable. L'approche habituelle pour la détection des ET est basée sur la similarité. Elle consiste à rechercher un alignement de séquences entre les séquences génomiques et les séquences nucléiques des ET connus (ex: BLASTN, RepeatMasker : http://repeatmasker.genome.washington.edu). Nous avons développé des outils d'annotations mieux adaptés à la détection des ET car tenant compte des difficultés suivantes :

- Dans un génome, les copies d'un ET sont très morcelées : chez *D. melanogaster* 50% des copies font moins de 15% de la longueur de l'élément complet. Les copies subissent de nombreux événements de délétion. De plus, il n'est pas rare que les copies de différents ET se retrouvent emboîtées les unes dans les autres. Tout ceci concourt à ce que l'alignement avec les séquences d'ET de référence nécessite de grands gaps. Les programmes généralement utilisés (BLAST, RepeatMasker, ….) s'interrompent dans ces gaps. Le résultat place alors dans des alignements indépendant chacun des fragments autour de ces grands gaps. On détecte alors autant d'ET différents, là où il n'en existe qu'un seul.
- Les séquences génomiques sont longues. Il n'est souvent pas possible de réaliser un alignement avec toute la séquence en une seule étape. Il faut alors fractionner celle-ci, aligner, et recalculer les coordonnées des séquences alignées dans un système de coordonnées génomiques.
- L'analyse d'un génome entier demande des comparaisons massives nécessitant des ressources informatiques importantes qu'il faut rationaliser.

Nous avons développé trois programmes nous permettant de travailler sur de telles séquences : BLASTER, MATCHER, GROUPER (Quesneville *et al,* 2003).

BLASTER

BLASTER est un programme développé en C++ au laboratoire. Il réalise une comparaison entre deux ensembles de séquences : une banque de séquences « query » face à une banque « subject ». Il peut également rechercher les séquences répétées quand les banques « query » et « subject » sont les mêmes. Quatre étapes sont nécessaires:

1. Découpage des longues séquences en fragments chevauchants afin de ne pas être limité par la longueur des séquences. Cette étape permet de travailler sur des génomes entiers.
2. Chaque séquence de la banque "query" est utilisée tour à tour pour une recherche par similitude dans la banque "subject" par l'un des programmes BLAST (BLASTN, TBLASTN, BLASTX, TBLASTX, BLASTP, MEGABLAST) (Altschul et al., 1990, 1997).
3. On élimine certains des alignements de séquences (HSP, pour High Scoring Pair, dans la terminologie de BLAST) : ceux dont la probabilité d'apparition par le hasard (mesurée d'après la E-value) est grande et/ou ceux dont le pourcentage d'identité est faible et/ou encore ceux dont la longueur est trop petite. Les seuils sont donnés par l'utilisateur.
4. L'étape 1 engendre des alignements chevauchants qui sont ré-assemblés à cette étape.

Les BLAST peuvent être distribués sur un cluster de machine à l'aide de la bibliothèque PVM (Parallel Virtual Machine). Les résultats obtenus sont traités soit par le programme MATCHER, soit par le programme GROUPER en fonction des opérations à réaliser ultérieurement.

MATCHER

Le programme MATCHER permet de reprendre les résultats de BLASTER. Son rôle est de positionner les « matchs » (les HSP) sur les séquences « query ». Lorsque MATCHER est confronté à des « matchs » se chevauchant sur une même région de la séquence « query », il sélectionne le match ayant le score d'alignement le plus élevé.

Les alignements contigus sur les séquences « query » et « subject » sont réunis par un algorithme d'alignement de fragments utilisant la programmation dynamique. En effet, si deux séquences homologues diffèrent par de longues insertions (ou délétions), les programmes BLAST reportent deux HSPs, un pour chaque région de part et d'autre de l'insertion/deletion. Les copies des ET subissent fréquemment de grandes délétions internes, c'est pour cela qu'ici MATCHER corrige ce défaut en réunissant les HSP contigus.

Nous utilisons un algorithme de chaînage en deux dimensions de paires de sous séquences non-chevauchantes utilisant la programmation dynamique (Gusfield 1997).

I. Un score est attribué à chaque connexion entre deux HSP. Celui-ci fait intervenir les scores d'alignement des deux HSP (résultats de BLAST) moins une pénalité de « gap » et une pénalité de « mismatch » telles que présentées par Chao *et al* (1995).

II. Le chaînage d'HSP retenu est celui pour lequel la somme des scores de connexion est la plus élevée. L'algorithme a été modifié pour produire le meilleur chaînage de type « alignement local », et non pas un chaînage de type « alignement global ». L'approche utilisée est similaire à celle de l'algorithme de Smith et Waterman, où on ne conserve que les chaînages donnant un score positif et une somme des scores maximum.

III. Le meilleur chaînage obtenu est sauvegardé, puis retiré avec ses HSP pour effectuer une nouvelle recherche. Ainsi, itérativement tous les chaînages sont recherchés.

Ces résultats étant difficilement lisibles tels quels, nous avons développé un outil permettant la transformation des fichiers de sorties de MATCHER en un format de fichier lisible par l'intermédiaire d'un éditeur d'annotation de séquences : le logiciel Apollo (Lewis *et al.* 2002). Apollo est un outil d'annotation des génomes qui permet d'éditer et de visualiser ces annotations. Il a été développé en collaboration entre le Berkeley Drosophila Genome Projet (BDGP) et l'Institut Sanger à Cambridge. Il permet d'explorer des annotations génomiques à plusieurs niveaux de détails graphiques à l'aide de zooms. Le format de fichier correspond à un ensemble de

spécifications en langage XML. Ce format appelé gameXML, est compatible avec bon nombre d'outils développés par BDGP.

GROUPER

GROUPER est un programme qui traite les résultats de BLASTER dans le but de réunir les alignements obtenus en groupes de séquences similaires. Deux types de regroupements sont construits par un algorithme de type « clustering » simple lien : les *groupes* et les *clusters*.

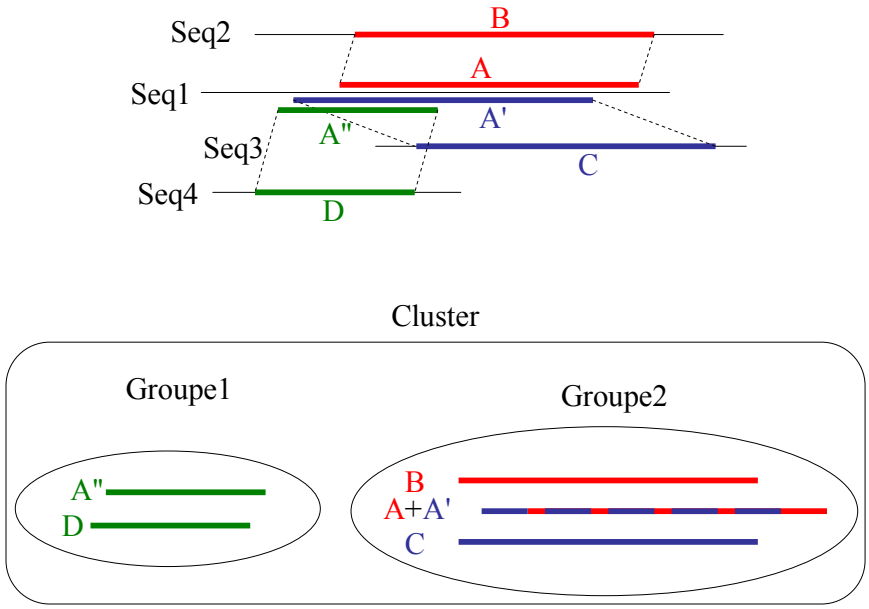

Figure 1: *Groupes* et *Clusters* obtenus par GROUPER

Les alignements obtenus par BLASTER sont examinés tour à tour. Soit *A1* et *A2* deux séquences d'un alignement *A*, et un *groupe G* de *n* séquences. Les séquences *A1* et *A2* font parti du *groupe G* si *A1* ou *A2* chevauche sur au moins 95% de sa longueur une des séquences du *groupe G* (elles partagent une région de même coordonnées sur une des séquences des banques « query » ou « subject »). Les séquences chevauchantes sont fondues en une seule les recouvrant toutes les deux. Si aucun *groupe* existant ne possède de séquence chevauchante, alors un nouveau *groupe* est créé avec *A1* et *A2*.

Les *groupes* sont réunis en *clusters* lorsqu'une des séquences d'un *groupe* chevauche l'une d'un autre *groupe* (Figure 1). Ici aucune limite sur la longueur du chevauchement n'est appliquée. On peut donc voir les *clusters* comme des *groupes* construits sans la contrainte de 95%.

Les résultats de GROUPER sont donc des *groupes* de séquences similaires sur leur plus grande longueur. A cause de la contrainte de longueur sur les *groupes*, une région donnée peut être présente dans plusieurs *groupes* contenus dans un même *cluster*. Par construction, deux séquences appartenant à deux *clusters* ne peuvent être alignées.

Recherche de nouveaux ET

L'alignement des séquences nucléotidiques (ex : par *BLASTN*), bien que précise, ne permet de détecter que des éléments connus, ou des séquences très proches de ceux-ci. Des séquences d'ET trop divergentes ou inconnues ne peuvent être détectées par la comparaison des séquences nucléiques. Nous recherchons ces éléments par deux approches différentes, l'une basée sur *TBLASTX*, l'autre sur un biais de composition nucléotidique.

Méthode BLASTER-TBLASTX

Le principe de la méthode est de détecter les séquences codantes similaires à celles de l'un des différents éléments connus. Cependant généralement la séquence codante des éléments de référence n'est pas connue, ou alors que très partiellement. En effet, on ne connaît souvent que des formes incomplètes des ET : leurs copies délétées. Notre stratégie a été de comparer les traductions dans les six phases des séquences nucléiques. Nous pouvons alors comparer des fragments de séquences codantes, et ainsi rechercher des similitudes au niveau protéique avec des séquences partiellement connues. Pour cela, nous utilisons *BLASTER* avec *TBLASTX*, puis *MATCHER* pour réunir les HSP adjacents. Le nombre de séquences d'ET identifiables est alors considérablement augmenté : cette méthode est très sensible (Quesneville *et al.* 2003).

Mais on ne détecte ici que les courtes régions conservées correspondant aux parties codantes. Il convient donc de récupérer les séquences flanquantes (généralement non-codantes) de ces régions et de reconstruire un élément consensus à partir des fragments de copies dispersés dans le génome.

- ✓ A l'aide de *BLASTER*, on réalise alors un *BLASTN* des séquences détectées par *BLASTER/TBLASTX* et *MATCHER*, étendues de 5000 paires de bases en 5' et en 3', contre les autres séquences étendues de la même manière, afin de détecter quelles sont les séquences qui possèdent entre elles des similitudes dans ces régions adjacentes. En effet, on s'attend à ce que les régions non-codantes des ET flanquant les régions identifiées par *TBLASTX*, se retrouvent aussi entre les différentes séquences étendues.

- ✓ Ensuite, les séquences vont être regroupées en *groupes* et *clusters* (par *GROUPER*) en fonction de leur similitude (au moins 90% d'identité).

- ✓ Un alignement multiple des séquences de chaque *cluster* permet d'obtenir une séquence consensus. Cependant, étant donné l'hétérogénéité des longueurs des séquences et leur nombre parfois important, l'alignement se fait en deux étapes sur les 20 plus longues séquences de chaque *cluster*:

    I. On réalise un alignement multiple de chaque *groupe*, puis on en déduit une séquence consensus en prenant la base majoritaire de chacune des colonnes de l'alignement.

    II. Pour chaque *cluster*, on réalise un alignement multiple des consensus obtenus sur les *groupes*. La séquence consensus est obtenue comme précédemment à partir de l'alignement multiple.

- ✓ Ces premiers consensus sont à nouveau comparés par *TBLASTX* aux ET connus afin de sélectionner parmi ces consensus ceux qui possèdent toujours des similitudes avec ces ET.

Biais de composition nucélotidique

L'approche *BLASTER-TBLASTX* reste encore très dépendante des séquences des ET connus. Nous développons en parallèle une approche basée sur les biais de compositions nucléotidiques qui devrait être moins dépendante d'une connaissance *a priori* des séquences d'ET (Andrieu O, Fiston A, Anxolabéhère D, Quesneville H, 2004).

Plusieurs observations ont montré un biais dans la composition nucléotidique des ET. La comparaison des séquences des ET avec celles des gènes chez *D. melanogaster* (Shields et Sharp 1989) a montré que les ET présentaient un plus grand nombre de codons se

terminant par les bases A ou T. L'étude de la composition en dinucléotides (Lerat *et al.* 2002) montre de plus, des fréquences de dinucléotidiques spécifiques du génome étudié. Ce biais peut être attribué au mode de réplication particulier des ET (retrotranscriptions, réparations des cassures double brin de l'ADN, …), mais aussi aux pressions de sélection particulières qui s'exercent sur leur séquence à cause de leur mobilité.

Notre méthode consiste dans un premier temps à estimer les paramètres d'un modèle probabiliste sur différents lots de séquences (séquences d'ET, gènes, ...). En travaillant selon un modèle probabiliste, on peut définir:

- S, une séquence représentée par une suite de symboles dans l'alphabet {A,T,G,C}
- L, l'ensemble des paramètres du modèle.

On peut alors calculer P(S/L) la probabilité conditionnelle d'avoir S sachant L. On peut ainsi calculer la probabilité des séquences selon les modèles suivants : les ET de classe I, de classe II, et les gènes. Il est ainsi possible de déterminer pour chaque séquence le meilleur modèle, c'est-à-dire le plus vraisemblable.

P(S/L) peut être calculé à l'aide d'une chaîne de Markov. Il s'agit d'un modèle probabiliste sur une suite finie d'observations où la probabilité d'une observation dépend des précédentes. On modélise ainsi la probabilité d'avoir un nucléotide à une certaine position *n* dans la séquence en fonction des positions *n-1, n-2,* ... . Cette dépendance est donnée par l'ordre du modèle.

- Une chaîne d'ordre 0 signifie que l'observation est indépendante des observations précédentes. (modèle de Bernoulli)
- Une chaîne d'ordre 1 signifie que l'observation dépend de l'observation précédente.
- Une chaîne d'ordre *n* signifie que l'observation dépend des *n* observations précédentes.

Cependant, une séquence biologique contient différents types de régions dont les plus connues sont les introns et les exons. Prises indépendamment, chacune de ces régions peut être considérée comme homogène en terme de composition nucléotidique ce qui

permettrait la construction d'une chaîne de Markov. Or ici, on ne connaît pas l'annotation en exon/intron des séquences. Il s'agit donc d'un modèle de Markov caché (HMM).

Les HMM sont des automates probabilistes permettant la modélisation de séquences tout en respectant leur hétérogénéité. Un automate probabiliste est une structure composée d'états, de transitions et d'un ensemble de probabilités de transition. Un symbole d'un alphabet fini, { A, T, G, C } dans notre cas, est généré à chaque état emprunté. Dans un HMM, chaque observation est le résultat d'une transition avec une probabilité notée a_{ij} qui est la probabilité de passer de l'état i vers l'état j suivi d'une émission avec une probabilité notée b_{jk} qui est la probabilité de l'observation k dans l'état j. Sans oublier les paramètres des états initiaux p_i qui représentent la probabilité d'être à l'état i initialement. Le modèle se représente donc sous la forme: $L=\{ p_i , a_{ij} , b_{jk} \}$ avec $i,j \leq n$ où n est le nombre d'états et $k \leq m$ où m est le nombre de symboles émis (ici m=4).

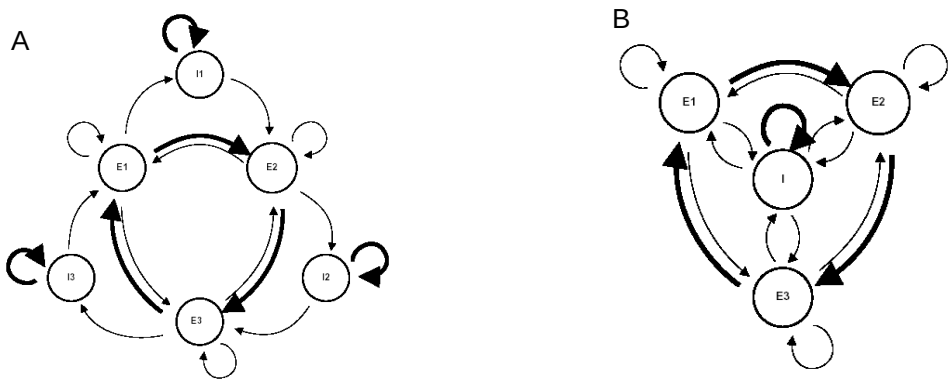

Figure 2 : Structures des HMM. Les états sont représentés par des cercles et les transitions par des flèches. L'épaisseur des flèches est proportionnelle à la probabilité de transition. **(A)** Modèle de gène : 3 états codants, 3 introniques **(B)** Modèle d'ET : 3 états codants, 1 non-codant.

- ✓ Le HMM utilisé pour les gènes contient 6 états dont 3 états exoniques (E1, E2, E3) et 3 états introniques (I1, I2, I3) (Figure 2A). Cette distinction des états est nécessaire pour prendre en compte les différences en composition nucléotidique entre les phases des régions codantes. Un nucléotide peut se trouver dans trois positions différentes dans un codon. Les trois états exoniques modélisent les positions dans le codon. De plus, dans ce modèle, si l'on passe de l'état E1 à l'état I1, en sortant de l'état intronique on est sûr de rester en phase si l'on a spécifié qu'on doit atteindre l'état E2. Ce HMM permet ainsi de prendre en compte la phase des séquences.
- ✓ Le HMM utilisé pour les ET, contrairement à celui des gène, ne contient que 4 états dont 3 états exoniques et un état non-codant (Figure 2B). Ce modèle prend également en compte les différences de composition nucléotidique des régions codantes, mais il est plus adapté aux ET. En effet, le HMM à 3 états introniques sont nécessaires pour garder la phase d'un exon à l'autre. Or les séquences d'ET à analyser peuvent avoir dégénérées et subit des délétions internes ou des mutations « frameshift ». La conservation de la phase d'un exon à l'autre n'est plus une caractéristique importante de la structure de ces séquences.

Pour rechercher de nouveaux ET, notre méthode nécessite une première étape de sélection des séquences candidates à tester par les HMM. Pour cela nous extrayons toutes les répétitions du génome que nous pouvons repérer par *BLASTER* et *GROUPER*. A partir du lot de séquences obtenu, nous retirons les régions contenant au moins un gène ou un ET reconnus par *BLASTN* ou par *TBLASTX*. Puis pour chaque séquence nous calculons sa probabilité selon les différents modèles *HMM*. Nous retenons pour chaque séquence le modèle pour lequel la séquence a une meilleure probabilité d'être observée.

3.1.2. Résultats

Annotation des ET dans les séquences génomiques

Dans un génome, les copies d'un ET sont très morcelées : Ainsi nous avons montré que chez *D. melanogaster* 50% des copies font moins de 15% de la longueur de l'élément complet (Quesneville *et al.*, 2003). Les copies subissent de nombreux événements de

délétion et se retrouvent fréquemment emboîtées les unes dans les autres (chez *D. melanogaster* 163 copies d'ET contiennent d'autres ET).

La comparaison de nos annotations du génome de *D. melanogaster* (release 3) avec celles réalisées par BDGP (Berkeley Drosophila Genome Project) disponibles sur Flybase (http://flybase.bio.indiana.edu/), montre que nos outils sont plus performants car beaucoup plus sensibles. De nombreux nouveaux petits fragments n'avaient pas été annotés par BDGP, mais aussi de nombreux éléments emboîtés n'avaient pas été détectés comme tels. La structure morcelée est mieux caractérisée.

Comparée aux résultats obtenus par RepeatMasker, BLASTER est plus précis pour déterminer les bornes exactes des ET, et retrouve beaucoup plus souvent la structure morcelée des copies des ET, en identifiant de façon plus complète ses fragments.

Nouveaux ET

L'approche utilisée précédemment ne permet de détecter que des éléments connus, ou des séquences proches de ceux-ci. Des séquences d'ET trop divergentes ou inconnues ne peuvent être détectées par la comparaison des séquences nucléiques.

Nous avons cherché ces nouveaux ET dans les génomes de *D. melanogaster* et *A. gambiae* par *BLASTER* avec *TBLASTX* (Quesneville *et al*. 2003). Cette méthode détecte les séquences codantes similaires aux différents éléments connus et actuellement répertoriés dans la banque de données « REPBASE UPDATE» dédiée aux éléments répétés. Nous avons pu mettre en évidence plus de 300 ET non encore détectés chez la Drosophile et 400 pour l'Anophèle (Quesneville *et al.* 2003).

A l'aide des modèles de Markov cachés, nous avons étudié les ET et les gènes de trois organismes: *D. melanogaster, C. elegans* et *A. thaliana* (Andrieu O, Fiston A, Anxolabéhère D, Quesneville H, 2004). On montre que les ET complets à intermédiaire ARN (éléments de classe I ou rétroéléments) et ceux à intermédiaire ADN (éléments de classe II ou transposons) ont une composition qui permet de les distinguer des gènes par des HMM. Cette approche permet également de montrer qu'on peut distinguer les ET de classe I de ceux de classe II. De plus, les biais nucléotidiques semblent être différents chez ces trois organismes. L'utilisation des HMM, qui prend en compte ce biais nucléotidique et l'hétérogénéité des séquences, représente donc bien un moyen de détection et d'annotation

des séquences d'ET. Nos tests chez *D. melanogaster* et *A. thaliana*, donnent des sensibilités de détection approchant les 70 % et des spécificités approchant les 100 %. Chez *C. elegans*, la sensibilité pour les ET de classe I est faible (autour de 30%) et la spécificité est proche 100%. La faible sensibilité de *C. elegans* s'explique en partie par le faible nombre d'ET de classe I chez cet organisme, ce qui n'a pas permis un bon apprentissage.

Pour l'étude d'un génome pour lequel peu ou pas d'ET sont connus, les HMM ne peuvent être appliqués tels quels étant donné qu'ils nécessitent une phase d'apprentissage sur les ET connus de l'organisme. Nous proposons donc pour ces génomes, une stratégie en deux étapes :

1. Utilisation de la méthode BLASTER-TBLASTX (c.f. Méthodes): Détection des séquences similaires aux ET connus répertoriés dans Repbase Update et construction de séquences consensus
2. Apprentissage des HMM sur les séquences consensus obtenues à l'étape 1
3. Détection sur le génome des régions dont le biais de composition est proche de celui des ET (appris lors de l'étape 2)

Nous avons évalué cette approche, que nous appellerons BLASTER-HMM par une analyse en aveugle du génome de *D. melanogaster,* ayant auparavant retirer tous les ET de Drosophile répertoriés dans la banque Repbase Update. Nous avons comparé ces résultats à ceux obtenus par deux autres programmes d'identification de nouveaux ET : RECON (Bao *et al*, 2002) identifiant les ET à partir des répétitions d'un génome et LTR_STRUC (McCarthy *et al.* 2003) recherchant les ET à LTR.

Nous avons comparé les régions détectées par ces 3 méthodes à celles que nous prenons comme références obtenues avec BLASTER et MATCHER lorsqu'on connaît les ET. Les 3 méthodes donnent des résultats similaires autour de 86% de recouvrement de l'annotation prise en référence. Elles diffèrent cependant par la fraction non-recouvrante qu'elles reconnaissent. Ces régions correspondent potentiellement à des ET encore inconnus chez *D. melanogaster*. BLASTER-HMM, RECON et LTR_STRUC donnent respectivement 80.5%, 27.5% et 40% de leurs prédictions comme non-chevauchantes à celles de référence. Les prédictions de ces 3 programmes prises conjointement recouvrent

96.5% des annotations de référence. Ce dernier résultat montre que ces 3 méthodes fournissent des résultats complémentaires.

Nous avons ensuite cherché à construire les séquences consensus des familles d'ET identifiés par alignement multiple des 20 plus longues copies de chaque famille. Nous avons comparé les séquences consensus obtenus aux séquences des ET connus. BLASTER-HMM, RECON et LTR_STRUC permettent de reconstruire respectivement 26, 45 et 17 copies complètes. Le reste correspond à des consensus de fragments des éléments et des consensus hybrides constitués de morceaux d'ET différents. RECON semble être la plus efficace, mais au prix d'un nombre de familles très élevés (1370 familles identifiées !) dont beaucoup correspondent en réalité à des fragments d'une même famille d'ET. On peut également remarquer que les 3 méthodes retrouvent un certains nombre ET différents, ce qui illustre à nouveaux leur complémentarité.

3.2. ET ET DUPLICATIONS SEGMENTALES

Nos outils d'analyse nous permettent d'avoir une image plus précise de l'organisation des ET dans un génome (en y intégrant en particulier les nouveaux ET). Par exemple, nous avons déjà pu constater que les copies étaient beaucoup plus fragmentées et dispersées que ce qui était connu (Kaminker *et al* 2002). Ceci est à prendre en considération pour étudier l'importance des recombinaisons ectopiques entre copies d'un ET.

Les séquences d'ADN répétées sont de formidables agents restructurant des génomes. Nous connaissons quatre grandes classes de séquences répétées: les répétitions en tandem, les pseudo-gènes, les éléments transposables (ET) et les duplications segmentales (DS).

3.2.1. Les duplications segmentales

Les DS sont des blocs dupliqués d'ADN génomique dont la taille est généralement comprise entre 1 et 200 kb. Elles contiennent souvent des séquences telles que des satellites et des séquences géniques avec intron et exon. Les DS se composent donc d'ADN apparemment banal, les rendant indétectables par leur structure interne. Ainsi, la majorité des DS précédemment identifiées l'ont été grâce à des analyses *in silico* recherchant des séquences répétés (Bailey *et al.*, 2001). L'une des caractéristiques importante des DS est la

forte similitude de ses copies (>95%), ce qui suggère des évènements de duplication récents.

Leur implication dans de nombreuses maladies humaines a été largement relatée dans la littérature. Ainsi 11 pathologies ont été décrites comme dues à des accident chromosomiques médiées par des DS comme par exemple la maladie de Charcot-Marie-Tooth, la neurofibromatose de type 1 et l'hémophilie A (Emanuel *et al.*, 2001). Actuellement, les études portant sur la distribution des DS concernent surtout l'espèce humaine. Ces recherches montrent qu'une large partie des DS sont situées dans les régions péricentromériques et subtélomériques (Horvath *et al.* 2000a et b).

Ces études ont permis de proposer un modèle de formation des DS qui repose sur la présence de diverses séquences (gènes, ET, ADN inter-génique) ayant « transposé » (Samonte *et al.* 2002). Le modèle propose que la formation de ces duplications suivrait deux étapes :

1. Dans un premier temps, des événements de « transposition » successifs et localisés de courtes séquences génomiques. On peut envisager que la localisation de ces transpositions soit due à certaines caractéristiques propres aux sites accepteurs, par exemple la composition nucléotidique de l'ADN ou la structure locale de la chromatine. Cette supposition repose sur l'observation de nombreuses duplications dans l'ADN centromérique et péricentromérique.
2. La deuxième étape consiste en l'évènement de duplication proprement dit. C'est-à-dire le transfert physique de la séquence génomique d'un locus à un autre. Ce transfert est réalisé par l'intermédiaire des séquences qui ont transposé lors de la première étape. En effet la présence de ces séquences répétées favorise les événements de recombinaison homologue non-allelique qui ont la propriété de dupliquer le matériel génétique compris entre ces petites séquences répétées.

A partir de ce modèle, nous avons testé si les séquences génomiques transposées au niveau des séquences accepteuses étaient généralement des ET. Cette hypothèse repose d'une part sur le fait que les ET possèdent les propriétés de mobilité par transposition, nécessaires pour la première étape de formation des duplications segmentales selon le modèle de Samonte *et al.* (2002). D'autre part, il a été montré que les ET pouvaient induire des recombinaisons ectopiques qui conduisaient à des remaniements chromosomiques

(Gray 2000). Enfin, les résultats obtenus précédemment au laboratoire (Quesneville *et al*. 2003) montrent que 86% des grandes séquences répétées possèdent un ET et que 32% en contiennent au moins deux.

3.2.2. Résultats

Nous avons réalisé cette étude des DS chez *D. melanogaster*. Nous avons réalisé une étude de la structure des DS et cherché l'implication des ET dans celles-ci. Pour identifier les DS, nous avons entrepris une recherche des séquences répétées de *D. melanogaster* avec BLASTER et GROUPER. La difficulté majeur de cette approche est l'impossibilité de différencier alors les DS des ET dans les séquences répétées. L'ambiguïté existante entre les duplications et les ET est due, d'une part à la taille de ces répétitions : de 100 pb à 200 kb pour les duplications et de 80 pb à 15 kb pour les ET, et d'autre part au fait que ces séquences répétées sont généralement soit des fragments d'ET soit des duplications de segments contenant eux-mêmes des ET. Ces deux caractéristiques propres à ce type de séquence entraînent, lors d'un alignement de séquences répétées, l'apparition de groupes de séquences alignées dont la taille est régulièrement comprise entre 100 pb et 15 kb ce qui les rendent non distinguables.

Nous avons donc utilisé pour cette étude notre annotation des ET réalisée avec nos outils. Les DS ont alors pu être identifiées à l'aide des coordonnées des ET préalablement détectés. Nous avons considéré comme DS toutes les séquences répétées ne correspondant pas à un ET. Ainsi, on conserve les fragments répétés même s'ils contiennent plusieurs ET. Par précaution supplémentaire nous n'avons conservé que les séquences qui étaient recouvertes à moins de 95% par des ET.

Ainsi, nous avons pu détecter 101 DS ayant entre 2 et 10 copies. Leurs caractéristiques sont les suivantes :

- Nous montrons, comme dans le génome humain, la présence d'une accumulation des DS au niveau des régions péricentromériques, mais pas au niveau des régions subtélomériques. De plus, nous avons détecté une sur-abondance des DS sur le chromosome 4. Ces résultats suggèrent une préférence des DS pour les régions de type hétérochromatiques.

- 88% des duplications sont intra-chromosomiques
- Plus de 75% des duplications ont un nombre de copies égal à deux
- Plus de 3/4 des copies de duplication partagent 100% d'identité de séquence
- 50% des copies de duplication se situent à moins de 7 kb de distance
- La taille maximale est de 50 kb.

Nous avons ensuite recherché le long des duplications, et plus particulièrement aux extrémités de celles-ci, des ET qui pourraient être responsables de l'événement de duplication. Le recouvrement moyen en ET des DS est de 38% et 58% des copies ont au moins un ET au niveau de l'un de ses points de cassure. Ainsi, nous avons proposé un modèle expliquant la structure et la dynamique des DS. Ce modèle propose que la formation des DS serait médiée par les ET, non pas comme le résultat de leur activité de transposition, mais en raison de leur caractère répété dans le génome. L'homologie de séquence entre les fragments de deux copies répétées pourrait lors de la réparation de cassures doubles brins de l'ADN par un mécanisme de conversion génique entre séquences homologues non-allèliques, produire des DS

Ce modèle propose que l'homologie de séquence des fragments d'ET, plutôt que leur activité de transposition, serait en partie responsable de la formation des DS grâce à, mécanisme employé. Le processus se déroule en cinq étapes:

1. Dans un premier temps la formation de la DS nécessite la présence de deux séquences (A et B) différentes à l'exception d'une petite région similaire commune au deux séquences (Figure 3A). Ces petites régions sont généralement des fragments d'ET qui peuvent se retrouver ainsi localisées à diverses régions du génome.
2. La séquence B subit ensuite une cassure doubles brins (DSB) à proximité du fragment en commun. Cette DSB est suivie par une dégradation exonucléasique des brins d'ADN de part et d'autre de la DSB (Figure 3B).
3. Le système de réparation des DSB est ensuite activé pour réparer la cassure. Ce mécanisme utilise alors l'homologie de séquences du fragment, anciennement ET, pour rechercher et trouver l'autre séquence (séquence A) d'ADN. Elle est alors utilisée comme matrice pour réparer l'un des brin de la séquence (B) suivant le modèle de

réparation des DSB proposé par Nassif et *al.* (1994). Cette réparation entraîne la formation d'une structure de type « bulle » sur la chromatide matrice de la réparation. C'est la progression de cette « bulle » qui permet aux polymérases de réparer le brin (Figure 3C).

4. La réparation d'après la chromatide de A est suivie par un décrochage dont la position est aléatoire (Figure 3D). Ce type de décrochage a déjà été observé chez *D. melanogaster* lors de la réparation des DSB induite par l'élément *P*. Le fragment ainsi libéré va alors rechercher une très courte région similaire sur l'autre extrémité du brin ayant subit la cassure (B), ce qui va permettre l'appariement des deux extrémité du brin réparé et l'initiation de la réparation du second brin de B en utilisant le brin déjà réparé comme matrice (Figure 3E).

5. La réparation de la deuxième chromatide se termine par la ligation des fragments réparés. Ce mécanisme a permis le transfert d'une partie de la séquence A à l'intérieur de la séquence B en impliquant uniquement un court fragment d'ET à l'une des extrémités (point de cassure) de la séquence transférée (Figure 3F).

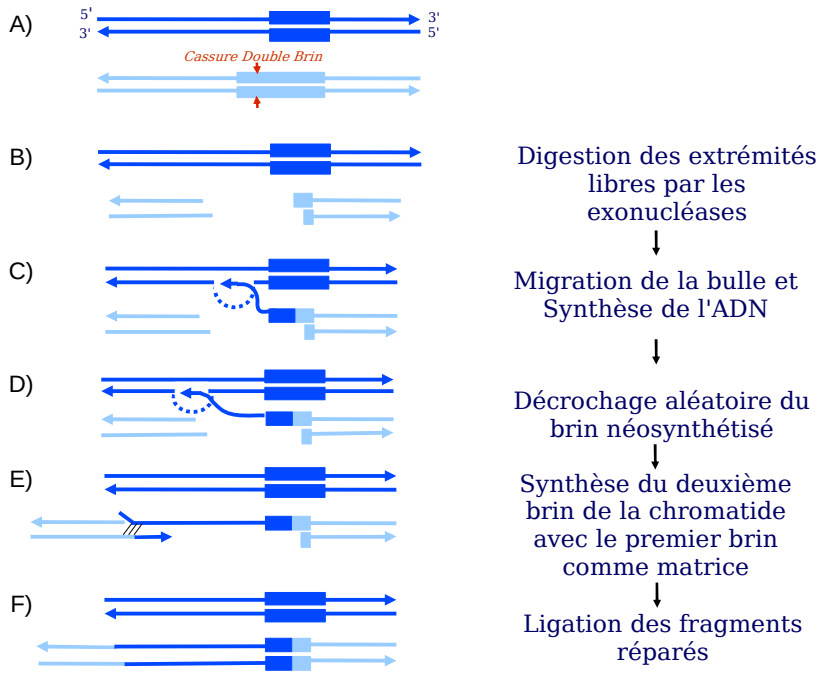

Figure 3: Modèle de formation des DS

Ce modèle n'implique pas l'utilisation des mécanismes de transposition des ET qui nécessitent généralement la présence de copies entières d'ET, que l'on retrouve rarement dans les DS étudiées.

Nous avons mis à l'épreuve ce modèle en cherchant d'abord à corréler la présence des DS à celle des ET.

- Nous avons comparé la distribution génomique des DS à celle des ET. Leurs densités sur les chromosomes sont significativement différentes. On peut remarquer une densité en DS particulièrement élevée sur le chromosome X par rapport à celle des ET.
- Par bootstrap, nous avons testé si l'occurrence élevée en ET était dû au hasard. Nous avons ré-échantillonné dans les régions génomiques où les DS avaient été identifiés, pour obtenir un lot de séquences témoins. Le recouvrement moyen en ET des DS observé (38%) n'est pas différent de ce qui est trouvé dans les séquences témoins. On retrouve également dans les séquences ré-échantillonnées, le chiffre observé de l'occurrence des ET aux extrémités des DS.

Ces résultats indiquent que les ET ne semblent pas impliqués dans la formation des DS, mais qu'ils ont la même affinité pour les régions hétérochromatiques. Ceci n'exclue pas qu'une minorité est été créées par des ET comme cela a déjà été rapporté dans la littérature (Gray, 2000).

Nous avons également testé le modèle de réparation en recherchant sur les séquences dupliquées les traces caractéristiques laissées par celui-ci. En particulier, nous avons cherché des séquences répétées en tandem de type microsatellite laissées par un décrochage prématuré du brin néo-synthétisé de son brin matrice suivit d'un raccrochage en amont. Plus de 70% des DS que nous avons identifié possèdent ce type de trace. Ce qui semble confirmer l'implication de ce mécanisme de réparation.

En conclusion, nous n'avons pu mettre en évidence un lien de cause à effet entre ET et DS. Il semble que la co-occurrence des DS et des ET soit liée au fait qu'ils cohabitent dans les mêmes régions génomiques. Cependant, il semble que le mécanisme de réparation homologue suivant le modèle SDSA (Nassif *et al*, 1994) puisse être impliqué dans la

formation de ces DS. Ces résultats nous amènent à penser que la première étape du modèle présenté n'implique plus des fragments d'ET, mais plus probablement de courtes séquences homologues permettant un accrochage du brin lésé en vu de sa réparation.

3.3. MODELISATION DE LA DYNAMIQUE DES GENOMES

3.3.1. Modélisation de la dynamique des ET

Nous avons étudié, par modélisation, l'origine et l'évolution des ET (Quesneville and Anxolabéhère, 2001). En effet, l'origine des ET est encore aujourd'hui très spéculative. De récents résultats suggèrent qu'ils puissent être à l'origine des rétrovirus, et seraient donc des vestiges d'un monde pré-biotique à ARN. D'autres résultats suggèrent que certains éléments puissent dériver de gènes endogènes (de leur hôte) possédant, par exemple, une activité endonucléase (recombinase, télomérase, enzyme de restriction, …*etc*.). Les modalités de leur évolution restent également tout aussi mystérieuses. En effet, les ET se répliquant plus souvent que des séquences immobiles, et mettant en œuvre pour leur transposition une machinerie de réplication différente, subissent un fort biais mutationnel imposant une évolution particulière de leurs séquences. D'autre part, les séquences étant présentes en grand nombre dans un individu, les pressions de sélection s'exerçant sur une copie d'un ET dépend de la nature et de la quantité des autres copies présentes chez l'individu qui porte celle-ci. Elles subissent donc une sélection de groupe doublée d'une sélection fréquence dépendante.

Méthode

Nous avons abordé ce sujet sous un angle *systémique.* Cette voie d'étude, issue des travaux théoriques récents sur les systèmes dits « *complexes* », recherche des « lois de l'organisation ». Elle consiste à rechercher les propriétés intrinsèques d'auto-organisation d'un système par l'exploration des interactions élémentaires permettant l'émergence d'un phénomène global. Ainsi, l'évolution des ET au sein de leurs hôtes est le résultat d'une sélection agissant à la fois sur les éléments et sur le génome de ces hôtes : il y a co-adaptation. Les stratégies adaptatives possibles sont conditionnées par leurs séquences en acides nucléiques et la dynamique génomique de leurs hôtes. Nous avons recherché les stratégies possibles contraintes par la structure d'un ET à ADN et les mécanismes agissant dessus. Nous avons recherché une description minimale d'un ET à ADN en terme de:

séquence, propriétés des gènes de l'hôte, propriétés de l'élément, mécanismes de transposition, *etc...* Nous avons ainsi identifié des caractéristiques générales nécessaires à leur émergence. Nous pouvons donc imaginer leurs structures minimales, et ainsi comprendre leur origine et leur évolution.

Modèle

Nous considérons les copies d'un ET sous la forme de séquences binaires insérées dans le génome d'individus constitués en populations. Quatre régions de cette séquence déterminent les caractéristiques fondamentales d'un élément (Figure 4). Deux régions définissent les propriétés de la transposase (enzyme responsable de la mobilité). L'une caractérise la séquence de son site cible et donc spécifie les copies qu'elle peut mobiliser, et l'autre, son efficacité à réaliser une transposition. Les deux autres régions situées aux extrémités de la séquence codent le site cible de la copie reconnu par la transposase, et donc spécifient quelles molécules peuvent la mobiliser. Les chromosomes des individus subissent des recombinaisons, et les séquences, des mutations et des délétions générant de nouveaux variants. La survie des individus étant affectée par la mobilité des éléments, le système évolue vers un équilibre. Ce modèle est implémenté sous la forme d'un programme de simulation orienté-objets. Nous pouvons ainsi explorer les propriétés des séquences sélectionnées, mais aussi leurs distributions au sein des génomes et des populations.

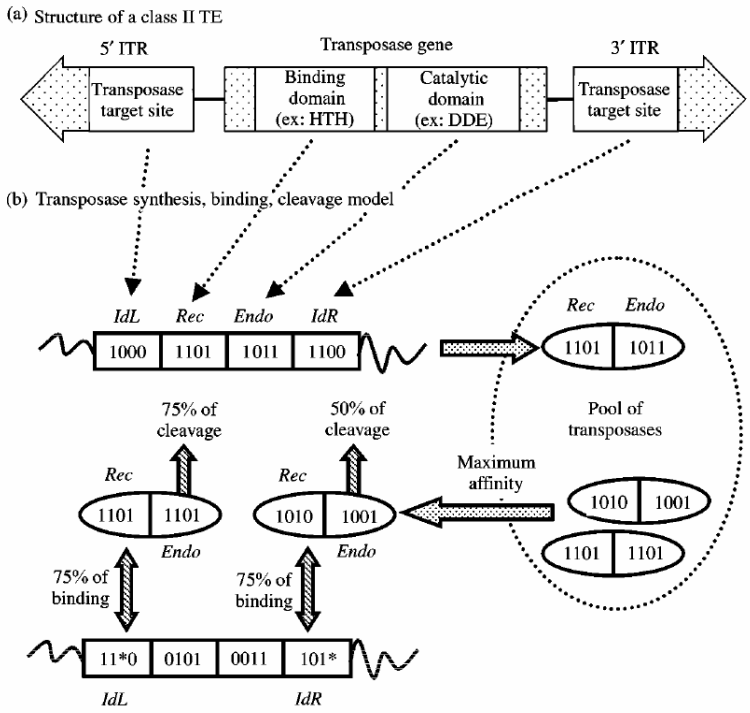

Figure 4 : (a) Représentation schématique d'un ET de classe II (b) Illustration du modèle. Synthèse de la transposase, reconnaissance des sites cibles et clivage. La probabilité de fixation est donnée par la proportion des symboles identiques - à chaque position - entre les domaines «Rec » (reconnaissance) et « IdL » (ou « IdR »), sites cibles de la transposase droit (ou gauche). Le symbole '*' remplace n'importe quelle symbole. La probabilité de clivage est donnée par la proportion de '1' dans le domaine « Endo » .

Résultats

Nos résultats montrent qu'un ET à ADN peut émerger à partir d'un gène unique et immobile capable de réaliser des coupures de l'ADN. Ce nouvel élément, lorsqu'il se multiplie par transposition, exerce une pression de sélection qui favorise l'apparition de nouvelles copies régulatrices limitant la mobilité de celui-ci. Ainsi, nous avons vu apparaître une régulation par compétition sur le site cible de la transposase, entre transposase active et transposase tronquée produite par des éléments délétés, ainsi qu'une régulation par titration de la transposase active par les éléments délétés et inactifs. Ces résultats sont en accord avec les données biologiques publiées. En effet, l'élément *P* est connu pour mettre en jeu de telles répressions par des éléments délétés. De plus, de nombreuses familles d'éléments comme *Hobo* pour *D. melanogaster, Tam3* pour *Anthirinium majus*, *Ac* pour le Maïs*, Mariner* pour *D. mauritiana* et de nombreuses autres espèces animales allant des insectes aux mammifères sont connues pour posséder des

éléments délétés. Ceux-ci pourraient donc réguler l'activité des éléments complets par des mécanismes identiques à ceux que nous venons de décrire. La régulation pourrait être due soit à une titration de la transposase par des éléments délétés inactifs, soit à une transposase tronquée. Celle-ci inhiberait la transposase fonctionnelle par compétition sur les sites cibles. Dans notre modèle, le mécanisme de transposition génère ces éléments délétés fortuitement, cette régulation n'est donc pas le résultat d'une évolution progressive de copies qui acquièrent progressivement des propriétés régulatrices, mais plutôt une conséquence indirecte du mécanisme de transposition et de la structure en quatre régions de l'élément que nous proposons. Elle pourrait donc être généralisable aux familles d'ET partageant le même mécanisme de transposition comme les familles *Ac* (Maïs) et *Tam3* (*Anthirinium majus*). Ce mode de régulation est le premier candidat en tant que mécanisme général contrôlant la transposition des ET à intermédiaire ADN.

De plus, les interactions entre l'hôte, les copies actives et les copies régulatrices provoquent un conflit génétique qui se résout par une évolution antagoniste des différentes régions de l'ET et des différents types de copies générés. Ces résultats montrent que l'évolution des ET peut être le résultat d'un conflit génétique qui induit un biais mutationnel variable le long de leurs séquences.

3.4. MODELISATION MULTI-NIVEAUX DES SYSTEMES BIOLOGIQUES ET REPRESENTATION DES CONNAISSANCES.

L'approche modélisatrice que nous avons utilisée a soulevé un certain nombre de questions d'ordre méthodologique. Ces questions nous ont amené à expliciter la démarche sous-jacente à notre approche. Ainsi, nous avons poursuivi notre réflexion par la caractérisation de son champ d'application et des perspectives qu'elle ouvre. Ce travail fait l'objet de trois publications de Quesneville et Anxolabéhère dans : LMO'97 (actes du colloque *« langages et modèles à objets »*), une publication sur le web disponible à *http://dynagen.ijm.jussieu.fr/equipes/bioinfo/papers/destobio.ps,* et un chapitre dans un livre sur la bioinformatique édité par Springer (voir Annexes). Voici un bref aperçu des idées qui y sont avancées.

Pour comprendre la dynamique des systèmes biologiques complexes, il est nécessaire d'intégrer leurs différents niveaux d'abstraction. En effet, chaque composant à un " niveau

d'abstraction ” donné (ex : la population), peut être lui-même décrit à l'aide de composants de niveau inférieur (ex : l'individu). Dans de tels systèmes, le comportement global ne se déduit pas de la juxtaposition des comportements de ses composants : on parle alors de systèmes non-linéaires. Cette non-linéarité les rend extrêmement difficiles à modéliser et à étudier. En effet, les modélisateurs doivent faire face aux difficultés liées à cet emboîtement : représenter la hiérarchie entre les niveaux et traduire les interactions entres les composants du système appartenant à des niveaux différents. De plus, chacun des niveaux d'abstraction est souvent lui-même objet d'étude de plusieurs domaines de la biologie. On est alors amené à synthétiser la connaissance d'un système, c'est-à-dire regrouper des connaissances provenant de différents domaines de la biologie. La nature de ces connaissances peut être différente. On peut en distinguer trois grands types : des structures qui caractérisent des relations physiques entre les entités d'un système (molécules, chromosomes, biotope,...), des processus qui décrivent des mécanismes exprimés sous la forme d'une suite d'événements (recombinaison génétique, comportements,...), et des flux qui spécifient des échanges continus d'informations (réactions enzymatiques, réseaux métaboliques, flux d'énergies dans un écosystème,...).

Pouvoir formaliser un système avec le souci de conserver la hiérarchie des niveaux d'emboîtement, traduire les interactions entres éléments appartenant à des niveaux d'abstraction différents, et regrouper des connaissances de nature hétérogène, ouvrent des perspectives nouvelles. En effet, le travail de modélisation ne se limite plus à prédire des comportements globaux du système modélisé. Mais par la formalisation, il permet aussi de transcrire les connaissances dans un langage commun permettant un partage de concepts et une communication entre biologistes experts de différents domaines. De plus, cette formalisation permet de tester si le comportement d'un niveau de description s'explique par ce qu'on connaît des niveaux inférieurs, et ainsi de tester la cohérence des connaissances entre les domaines de la biologie.

Le paradigme « objet » offre un formalisme donnant les outils pour cette nouvelle approche. Nos travaux de modélisations (Quesneville et Anxolabéhère 1997a et b, 1998, 2001) illustrent leur mise en œuvre pour la modélisation de systèmes génétiques complexes. Les travaux de modélisation qui seront engagés développeront cette réflexion méthodologique.

3.5. ETUDE DES METHODES D'IDENTIFICATION DES FACTEURS DE SUSCEPTIBILITE IMPLIQUES DANS LES TRAITS MULTIFACTORIELS.

(Collaboration avec le laboratoire : Génétique épidémiologique et structure des populations humaines- INSERM 535 : Directeur : Françoise Clerget-Darpoux)

La stratégie de modélisation orientée-objets que nous avions développée a trouvé un autre champ d'application en génétique épidémiologique. Si les stratégies d'étude des traits monogéniques sont aujourd'hui bien établies, ce n'est pas le cas pour les traits multifactoriels. Les traits multifactoriels sont des caractères complexes déterminés par l'interaction d'un ensemble de facteurs génétiques et environnementaux. Les facteurs génétiques impliqués dans ces traits ne sont ni nécessaires ni suffisants à leur expression. De nombreuses maladies relèvent de ce type de déterminisme. Différentes méthodes ont été proposées pour identifier le rôle d'un facteur génétique impliqué dans un trait multifactoriel. On distingue deux catégories : les études d'associations et les études de liaisons. Elles reposent sur un certain nombre d'hypothèses quant à la population dont sont issus les individus étudiés : grande population, absence de structuration, de stratification, unions au hasard (on parle de panmixie), consanguinité, *etc*... Parallèlement au développement de ces méthodes, il est donc important d'étudier leur robustesse face à ces hypothèses, mais aussi de rechercher des stratégies d'étude adaptées à la nature des traits étudiés exploitant au mieux les propriétés de certaines populations (ex : petites populations consanguines, effets fondateurs, mariages entre apparentés, *etc*...).

3.5.1. Outils de simulation en génétique des populations

Dans cette perspective, nous avons étendu les capacités de simulation de notre simulateur dédié aux éléments transposable. Grâce à l'approche orientée-objets, il a été possible en quelques mois de construire le nouveau simulateur, appelé GENOOM, étendu à des populations spatialement distribuées, à effectifs variables, avec un mélange de classes d'éléments génétiques (microsatellites, marqueurs RFLP, gènes, QTL, …), une stérilité liée à certains phénotypes, des croisements entre apparentés, et des migrations. Cette évolution rapide illustre la notion de "réutilisabilité" du paradigme objet. Cet outil est actuellement utilisé à l'unité INSERM 535 pour tester diverses méthodes d'analyses des populations

humaines. Ainsi, GENOOM permet par simulation de déterminer la robustesse et la puissance des statistiques utilisées. Il permet également de rechercher des stratégies d'étude performantes, adaptées aux caractéristiques du trait étudié (nombre de locus, pénétrances, phénocopies, *etc*...), et en particulier de rechercher les caractéristiques optimum de population (niveau de consanguinité, population à effet fondateur, valeurs de déséquilibres gamétique, *etc*.). Un tel travail a donné lieu à une thèse (Bourgain 2001, thèse de l'université Paris XI), et plusieurs publications (Bourgain *et al*. 2000, 2001a, et b, 2002).

3.5.2. Le TTS (Triangle Test Statistic)

Les méthodes des paires de germains (couples frère-frère, frère-sœur, ou sœur-sœur) atteints sont aujourd'hui devenu un outil standard d'analyses de liaison pour les maladies multifactorielles. Le principe de ces méthodes est de compter le nombre d'allèles marqueurs identiques par descendance partagé par les germains, et de tester si ce nombre est différent de celui attendu sous l'hypothèse d'une ségrégation indépendante des marqueurs et de la maladie. Les proportions des paires de germains partageant 2, 1, ou 0 allèles marqueurs parentaux identiques par descendance, soient respectivement Z_2, Z_1 et Z_0, subissent les contraintes suivantes : $Z_0+Z_1+Z_2=1$ et $2\ Z_0 \leq Z_1 \leq 0{,}5$. Ces contraintes qui restreignent ces valeurs à appartenir à un triangle, ne tiennent plus lorsque les deux germains diffèrent par un facteur qui modifie leur probabilité d'être atteint. Ce facteur peut être une caractéristique du trait lui-même (ex : la sévérité), ou un facteur environnemental. Dans ces conditions la procédure classique du test de la liaison génétique perd en puissance de façon importante. Nous avons donc proposé une statistique, le TTS (Triangle Test Statistic) qui teste sur ce type de données à la fois la liaison et une hétérogénéïté due au facteur (Quesneville *et al.* 1999, Dizier *et al.* 2000).

Cette approche a été appliquée sur des maladies connues pour posséder une telle hétérogénéïté. Elle a permis d'identifier de nouveaux marqueurs liés à des locus de susceptibilité de ces pathologies (Dizier *et al.* 2001a et b, 2003) .

4. PERSPECTIVES ET PROJET

Notre projet porte sur l'étude de l'impact des ET sur les génomes. Ce projet s'articule autour de trois axes complémentaires.

1) La détection et l'identification des ET dans les séquences génomiques. L'étude de l'organisation des ET dans un génome se situe en amont de tout notre travail. Décrire précisément la distribution et la structure des ET au sein d'un génome est un préalable indispensable à la compréhension de sa dynamique et de ses interactions avec les ET.

2) Une modélisation de la dynamique évolutive des génomes sous l'effet des ET. Il a été montré que les ET étaient à l'origine de nombreux remaniements chromosomiques tels que des délétions, des inversions, des duplications et des translocations. Véritable moteur de la fluidité des génomes, la dynamique de ces réarrangements sera étudiée par modélisation et simulation. Notre travail d'annotation des ET dans les génomes fournira les données nécessaires à la validation de nos modèles, et réciproquement fera émerger des questions à résoudre par nos modélisations.

3) Impact des produits des ET sur le génome de leur hôte. Les produits codés par les ET ont des propriétés endonucléases, recombinases, polymérases, *etc* … Ces activités ne sont pas neutres pour le génome hôte, indépendamment de leurs actions sur les ET eux-mêmes. Nous souhaitons nous intéresser à leurs effets sur le génome et étudier comment elles peuvent être « recrutées » par celui-ci.

La pluri-disciplinarité de ce projet est essentielle. Les interactions entre ces 3 axes sont voulues comme un élément moteur de la dynamique scientifique de l'équipe.

4.1. ANNOTATION DES ET

La comparaison de nos annotations du génome de *D. melanogaster* (release 3) avec celles réalisées par BDGP (Berkeley Drosophila Genome Project) montre que nos outils sont plus performants car beaucoup plus sensibles. De plus, notre première étude réalisée sur *D. melanogaster* et *A. gambiae* (Quesneville *et al.* 2003) a montré qu'il restait encore beaucoup d'ET à décrire, même pour un organisme aussi bien connu que celui de la Drosophile. Nous n'avons pour l'instant détecté que de courtes régions conservées correspondant aux parties codantes d'ET non encore annotés dans ces génomes. Il convient donc de récupérer les séquences généralement non-codantes, flanquant ces régions et de reconstruire un élément consensus à partir des fragments de copies dispersées dans le

génome. Nous envisageons de construire, pour ces nouveaux ET, les séquences consensus par les méthodes présentées. Ces consensus représentent alors une approximation de l'état ancestral de ces copies et donc de l'élément actif.

Ces méthodes de recherche de nouveaux ET sont particulièrement utiles lorsqu'on étudie un organisme pour lequel peu d'ET sont connus comme pour *A. gambiae.* Ce sont les seules approches qui aujourd'hui permettraient de réaliser un inventaire correct des ET pour les génomes dont la composition en séquences répétées est mal connue. Il est important de souligner que l'ensemble des eucaryotes qui sont en cours de séquençage aujourd'hui sont dans cette catégorie. Il est donc essentiel de disposer d'outils et de méthodes capables de réaliser cet inventaire.

Forts de notre expérience et de ces résultats, nous avons engagé plusieurs collaborations internationales

1. Une première collaboration internationale a été engagée avec Michael Ashburner et Casey Bergman (University of Cambridge) afin de ré-annoter les ET pour la release 4 du génome de *D. melanogaster*. L'annotation est à l'heure actuelle pratiquement terminée.
2. Deux nouveaux génomes de Drosophiles sont en cours de séquençage : *D. yakuba* séquencé avec une couverture de 8X, ainsi que 8 souches de *D. simulans* correspondant chacune à une population différente séquencée, une avec une couverture de 2X, les autres en 1X. Nous avons engagé une nouvelle collaboration avec Michael Ashburner et Casey Bergman pour l'annotation des ET de ces Drosophiles.
3. Les releases 1 à 4 du génome de *D. melanogaster* ne concernent que sa séquence euchromatique. Récemment les régions hétérochromatiques ont été également partiellement séquencées (Hoskins *et al.*, 2003). Leur analyse montre, comme attendu, une forte densité des ET dans ces régions : 52% de la séquence est composée d'ET (3.86% pour l'euchromatine). Les séquences identifiées sont très fragmentées et semblent souvent générées par l'insertion d'éléments les uns dans les autres, formant alors des structures imbriquées complexes. Le séquençage de l'hétérochromatine se poursuit, et nous avons pris contact avec Gary Karpen (Lawrence Berkeley National Laboratory in Berkeley) afin d'annoter plus complètement celle-ci. Nos outils

devraient se montrer particulièrement efficaces sur ces cas complexes. Il est intéressant de noter ici que Hoskins *et al.* (2003) mentionnent qu'un certain nombre d'ORF présentent des similarités avec des transposases. Ceci suggère que de nouveaux ET restent à décrire également dans ces régions. Nos outils d'identification de nouveaux ET devraient nous permettre de les caractériser.

4. Parallèlement, une seconde collaboration a été démarrée avec le consortium responsable du séquençage et de l'annotation du moustique *Anopheles gambiae*. Le but étant également d'annoter avec nos outils les ET du génome de l'Anophèle. Cette collaboration implique en particulier Charles Roth (Institut Pasteur, France), Paul Brey (Institut Pasteur, France), Zhijian Tu (Virginia Polytechnic Institut and State University, USA), Frank Collins (University of Notre Dame, USA), et l'équipe d'Ensembl (EMBL-EBI et Sanger Institut, Royaume Uni). Les résultats de ces annotations devraient être disponibles à partir du site Ensembl dédié au moustique (http://www.ensembl.org/Anopheles_gambiae/).

Ces projets nécessitent de développer un « pipeline » d'annotation dédié aux ET. Un tel « pipeline » comprend :

1. Un ensemble de programmes capables de détecter, d'annoter, et visualiser les séquences des ET. Parmi ces programmes, on trouvera les outils que nous avons développé pour annoter et rechercher de nouveaux ET (*BLASTER ,MATCHER, GROUPER, HMM*) mais aussi les autres programmes disponibles tels que : *RepeatMasker, Censor,* et *RECON*. Les résultats de tous ces programmes seront analysés et confrontés automatiquement afin de produire l'annotation la plus complète possible. Celle-ci pourra être visualisée conjointement avec les prédictions des différents programmes grâce au logiciel d'annotation Apollo.
2. Une base de données capable de stocker les séquences et les annotations des génomes disponibles, mais aussi les données générées par les programmes d'annotation.
3. Des programmes permettant d'enchaîner et distribuer les tâches sur un cluster d'ordinateurs, et de contrôler leur bon déroulement.

4.2. IMPACT DES ET SUR LE GENOME

4.2.1. ET et remaniements chromosomiques

Il a été montré que les ET transposables pouvaient être responsables de réarrangements chromosomiques tels que des délétions, des duplications, des inversions, ou encore des translocations (Gray, 2000). Plusieurs mécanismes ont été proposés :

1. Les ET agiraient comme des séquences répétées et dispersées, pouvant induire par leur identité, des recombinaisons ectopiques générant ces aberrations chromosomiques.
2. Les ET à intermédiaire ADN transposent généralement par un mécanisme de type couper-coller, grâce à une transposase. Parfois, celle-ci fait des erreurs dans les coupures double brin de l'ADN. Au lieu de cliver aux deux extrémités de la séquence de l'ET en *cis* sur la même chromatide, elle coupe en *trans,* chaque coté sur une chromatide différente. Il en résulte ce qu'on appelle une transposition alternative qui conduit après transposition, à des remaniements chromosomiques tels que des délétions, des duplications, des inversions, ou encore des translocations.
3. Deux transposons à intermédiaires ADN d'une même famille, situés proches l'un de l'autre, peuvent transposer comme un seul élément, emportant la séquence génomique qui les sépare. On parle alors de transposons composites chez les procaryotes ou de macro-transposons chez les eucaryotes. Ces événements produisent exclusivement des duplications.

Pour expliquer certaines duplications segmentales que nous avons observées chez *D. melanogaster*, nous avons proposé un modèle de duplications. Dans ce modèle, celles-ci sont induites par de courtes régions homologues qui perturbent le système de réparation des cassures de l'ADN de l'hôte selon le modèle « SDSA » (*Synthesis Dependant Strand Annealing*, Nassif *et al*, 1994). Les séquences génomiques des 4 Drosophilidés séquencés ouvrent de nouvelles perspectives. En recherchant les DS dans les régions synthéniques de *D. melanogaster* avec *D. simulans* et *D. yakuba*, nous allons pouvoir déterminer le sens de ces duplications et connaître le brin néo-formé. Ainsi nous pourrons tester si les traces du

SDSA que nous avons observé sont bien situées sur le nouveau brin. Nous pourrons également dater les DS et ainsi retracer leur histoire dans ce groupe d'espèce.

De plus, en étudiant les régions synthéniques de ces génomes, il sera possible de détecter les autres évènements qui passent inaperçus lorsqu'on ne dispose que d'un seul génome : les évènements de délétions, inversions, et translocations. Le séquençage de *D. yakuba* et plus particulièrement les séquences de 8 populations de *D. simulans*, ouvrent la voie à ce type d'étude. Etant impliqués dans l'annotation des ET de ces génomes, nous seront tout particulièrement bien placés pour étudier ces remaniements chromosomiques et l'éventuel impact des ET. Ces deux nouveaux génomes nous permettront d'aborder cette question à différentes échelles de temps : (i) au niveau inter-populations en étudiant les 8 populations de *D. simulans*, et (ii) à 3 niveaux inter-spécifiques en comparant *D. melanogaster* à *D. simulans* (distants de 3 Ma), *D. melanogaster* à *D. yakuba* (distants de 10 Ma), et *D. melanogaster* à *D. pseudoobscura* (distants de 50 Ma). Nous chercherons à déterminer le rôle et l'importance des ET dans ces remaniements.

Enfin, les réarrangements chromosomiques pouvant être délétères, il a été proposé qu'ils puissent être contre-sélectionnés, ainsi que les ET qui les auraient induits. Ces évènements agiraient alors comme une force réprimant l'invasion des ET dans une espèce. Cependant, ces réarrangements sont aussi les agents principaux de la fluidité des génomes et de leur évolution. Réarrangements délétères ou générant de la variabilité, l'impact des ET apparaît ici comme dual. Cette opposition est source de dynamiques évolutives tout à fait particulières. Nous souhaitons étendre notre modèle de dynamique des ET implémenté dans GENOOM pour intégrer les recombinaisons ectopiques, mais aussi les transpositions alternatives et notre modèle de duplication, afin d'étudier la dynamique résultant de la dualité des effets des ET sur la valeur sélective des individus. Quels sont les mécanismes et les évènements les moins délétères ? Quels sont les réarrangements les plus fréquents ? Comment un réarrangement apparu chez un individu, peut se fixer dans une population puis une espèce ?

4.2.2. Impact des produits des ET sur le génome

Nous postulons que les interactions entre ET et gènes permettent la mise en place de nouveaux réseaux métaboliques. De nombreux exemples montrent que les régions régulatrices des ET et en particulier celles qui fixent la transposase, peuvent agir en *cis* sur

l'expression spatio-temporelle des gènes situés à proximité. Les copies d'un ET peuvent donc mettre sous un même contrôle génétique un ensemble de gènes. L'ET, devient alors responsable de la coordination de leur expression, et met ainsi en place un nouveau réseau de régulation. Ces séquences d'ET peuvent alors être "récupérées" par le génome de l'hôte et être à l'origine de l'émergence de nouveautés génétiques. C'est un des processus de domestication moléculaire des ET. Nous voulons étudier la mise en place de nouveaux réseaux de régulation et rechercher les modules de régulation dérivés d'ET qui coordonnent l'expression de gènes participant à un même réseau de régulation.

Nous avons choisi pour modèle d'étude l'ET *P* de *D. melanogaster*. Notre choix est motivé d'une part par l'observation d'interactions entre les produits de l'élément *P* et des gènes de drosophiles, et d'autre part le succès évolutif d'une partie de sa région codante au sein d'un taxon de plusieurs dizaines d'espèces (Nouaud D. *et al,*. *Mol. Biol. Evol.* 1997,1999, 2003).

La force de ce projet repose sur le couplage étroit d'une analyse bioinformatique et d'une approche expérimentale. Il consiste à (i) caractériser les motifs nucléotidiques susceptibles de fixer les produits de l'élément *P*, (ii) les détecter dans la séquence génomique (insertion d'élément *P*, vestiges d'anciens ET, séquences fortuites), (iii) tester si l'expression des gènes situés à leur proximité est sensible à leur produits, et (iv) rechercher par puces à ADN, les gènes différemment exprimés en présence des produits de l'élément *P*.

Présentation du modèle biologique

Nous possédons une longue expérience des ET, et plus particulièrement de l'élément *P* des *Drosophilidés*. Pour cette étude, nous nous intéresserons à la famille d'éléments *P* présente dans le génome de *D. melanogaster*. Ce modèle biologique offre des avantages exceptionnels pour ce type d'étude.

Drosophila melanogaster

Le génome de *D. melanogaster* est aujourd'hui entièrement séquencé (sauf toutefois certaines régions hétérochromatiques). La séquence est mise à disposition depuis Avril 2000 par BDGP (Berkley Drosophila Genome Project). De plus, *D. melanogaster* est un

organisme de choix pour les études génétiques tant la panoplie des outils moléculaires et génétiques est grande.

***L'élément transposable* P**

L'élément transposable *P* est un des ET les mieux caractérisé tant du point de vue moléculaire, que génétique, et de son évolution. D'autre part, l'élément *P* ayant envahi le génome des populations naturelles de *D. melanogaster* au cours des années 1950 suite à un transfert horizontal à partir d'une espèce distante de 60 millions d'années, il existe dans les collections de souches, des lignées dont le génome est dépourvu de cet élément (celles collectées avant le début de l'invasion) et des lignées dont le génome possède des éléments *P*.

La transposition massive des éléments *P* est responsable d'un ensemble d'anomalies génétiques (stérilité, mutations, distorsion de ségrégation…). Celles-ci se produisent dans la lignée germinale des individus issus d'un croisement entre un mâle porteur d'éléments *P* (type P) avec une femelle dépourvue d'éléments (type M). L'absence quasi totale de ce syndrome dans la première génération du croisement réciproque (femelle P x mâle M) ainsi que dans les croisements P x P est due à un « état cytoplasmique » transmis maternellement réprimant la transposition des éléments *P* : le cytotype P.

La structure de l'élément *P* comporte quatre exons (Figure 5). Ils forment un transcrit monocistronique qui, par le jeu d'un épissage alternatif, code deux produits : les quatre exons codent la transposase tandis que les trois premiers exons seuls codent une transposase tronquée, répresseur de la transposition. L'élément *P* existant sous deux formes structurales, les éléments pleine longueur, autonomes pour leur transposition, et les éléments présentant des délétions internes non-autonomes mais pouvant être mobilisés en *trans* si des éléments *P* autonomes, sont présents dans le génome. Les séquences de l'élément *cis*-nécessaires à la transposition sont restreintes aux 138 premières paires de bases en 5' et les 216 dernières en 3'.

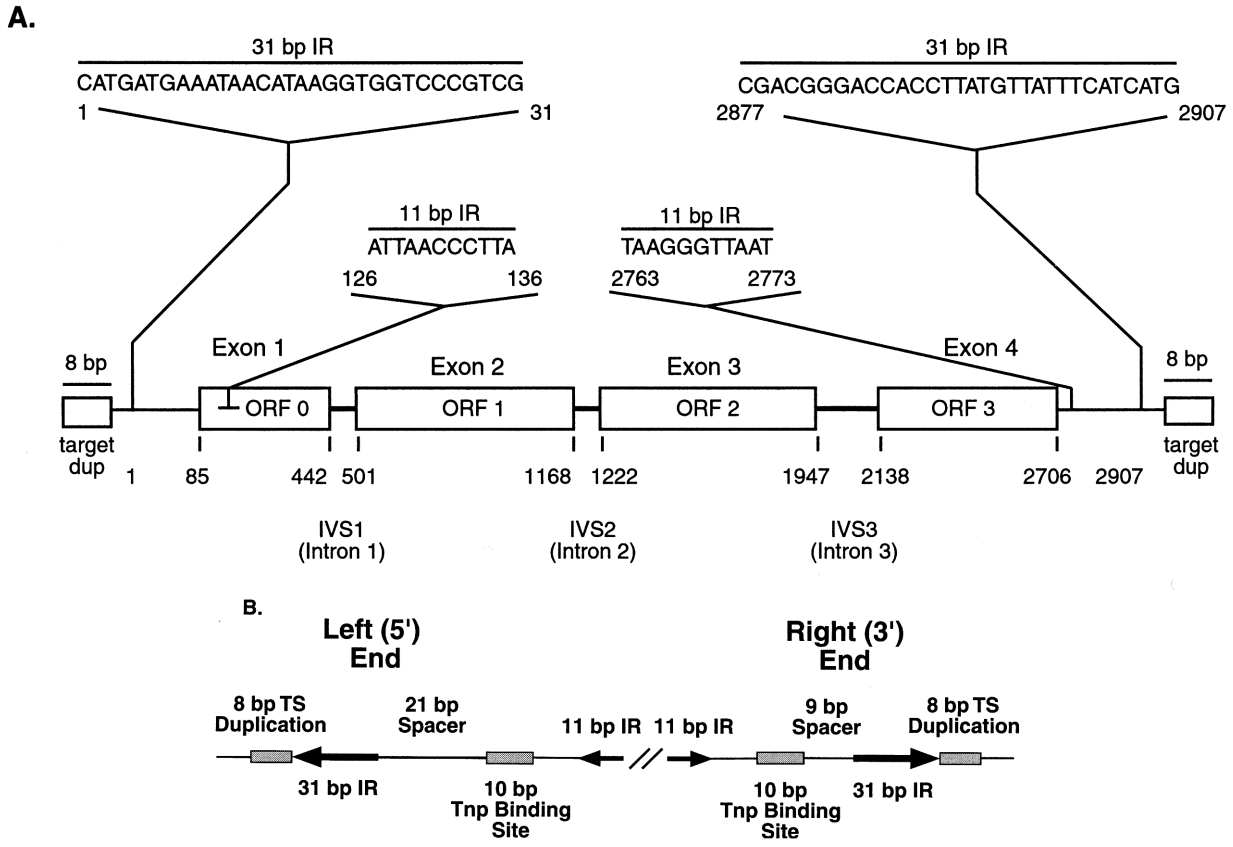

Figure 5 : (A) Caractéristiques de la séquence de l'élément *P*. (B) Eléments *cis*-actifs de la transposition de l'élément *P.* (D'après Rio 2001, *P* transposable elements in *Drosophila melanogaster*, in Mobile DNA II, Ed. Craig et al. ASM Press, Washington)

La transposase et le répresseur codés par l'élément *P* présentent, dans leur région commune, deux types de structures super-secondaires : un motif hélice-tour-hélice (H-T-H) et trois motifs « leucine zipper » (LZ). Ce motif H-T-H connu pour intervenir dans les interactions ADN-protéine site spécifique, pourrait correspondre ici à un domaine de liaison de la transposase à l'ADN. De plus, dans de nombreuses protéines de liaison à l'ADN, les motifs LZ sont impliqués dans la formation d'homo- ou d'hétérodimères protéiques. Ceci pourrait indiquer une aptitude de la transposase à rapprocher dans l'espace des sites de fixation distants en formant une boucle, mécanisme qui pourrait être impliqué dans la transposition de l'élément.

Interactions des produits de l'élément* P *avec des gènes

Il existe plusieurs exemples pour lesquels un gène s'exprime différemment en présence ou en absence des produits de l'élément *P*. L'exemple peut-être le plus démonstratif est leur effet répresseur sur l'expression des transgènes *P* insérés dans le génome de certaines lignées de drosophiles. L'expression de ces transgènes est souvent dépendante de la présence ou de l'absence des produits de *P*. Ainsi, l'insertion d'un transgène *P-lac(Z)* (fusion entre le début de l'élément *P* et le gène de la *beta-galactosidase* d'*E. coli*) est différemment exprimé en l'absence ou en présence des produits de l'élément *P*. Dans ce dernier contexte, les produits de *P* se fixent sur la séquence *P* du transgène et répriment fortement sa transcription. De plus, ce phénomène est fortement dépendant du site d'insertion du transgène : la nature des séquences environnantes exerce une action non encore élucidée sur la fixation des produits de l'élément *P*.

Ainsi nous postulons que la présence des protéines codées par les éléments *P* "perturbe" le niveau de l'expression de nombreux gènes, même si ceux-ci ne sont pas à proximité immédiate d'insertions d'éléments *P*. Des motifs nucléotidiques présentant une identité assez forte avec les séquences canoniques de fixation de la transposase et du répresseur pourraient constituer, *via* la fixation de ces protéines, des sites enhancer ou silencer modulant l'expression des gènes à proximité. Ainsi, la présence des produits de l'élément *P* aurait une action régulatrice sur de nombreux gènes indépendamment des mutations provoquées par l'insertion des éléments *P*.

Cette hypothèse pourrait rendre compte du succès évolutif de la domestication d'une partie de la séquence codante d'un élément *P* au cours de la différenciation d'un sous-groupe d'espèces de drosophiles: le sous-groupe *montium.* En effet, dans le génome de ces espèces, Nouaud *et al* (1997,1999, 2003) ont mis en évidence une domestication moléculaire d'un élément *P*. Ce néogène *P* ne comporte que les trois premiers exons de l'élément *P* canonique et a perdu toute mobilité à la suite d'une délétion terminale de la région 3' comprenant la totalité du quatrième exon (spécifique de la transposase). Ce néogène a conservé la capacité de coder une protéine "répresseur-like" présentant une similitude de 70% avec le répresseur de l'élément *P* canonique. Paradoxalement, il n'existe pas dans le génome de ces espèces d'éléments *P* mobiles de la même sous famille, une fonction répresseur n'est donc pas requise. Nous postulons que la capacité codante de ce

néogène a été maintenue parce que ce dernier subit une contrainte sélective positive. Celle-ci pourrait être due à la mise en place d'un réseau de régulation concernant des gènes dont l'expression est modulée par la protéine codée par le néogène *P*. L'émergence de ce réseau de régulation serait la conséquence de la labilité de la reconnaissance des motifs nucléotidiques par les protéines de l'élément *P* sur lesquels elles peuvent se fixer.

Caractérisation des motifs nucléotidiques susceptibles de fixer les produits de l'élément P

La transposase est une protéine de liaison à l'ADN site-spécifique, qui reconnaît deux séquences internes proches des extrémités de 31 pb répétées inversées de l'élément *P*. Elle se fixe en 5' sur une région de 21pb couvrant les nucléotides 48 à 68 et en 3', sur une région de 17 pb correspondant aux nucléotides 2855 à 2871. A partir de ces deux sites, une séquence consensus de liaison à la transposase a été établie : il s'agit d'une séquence de 10 pb, riche en AT : du coté 5' le consensus est AT(A/C)CAATTAA, il est en orientation inverse du coté 3' de l'élément. Il est très probable que la protéine répresseur se fixe sur ces mêmes séquences.

D'autre part des cas de mobilisations croisées entre des éléments *P* appartenant à des espèces de drosophiles différentes ont été rapportés. Le motif consensus décrit précédemment n'étant que partiellement présent dans les séquences mobilisées, nous supposons que celui-ci est moins strict. Nous allons donc chercher à l'affiner en le recherchant parmi l'ensemble des séquences des éléments *P* de drosophilidés connus pour avoir été mobilisées par la transposase de l'élément *P* de *D. melanogaster*.

La caractérisation des motifs nucléotidiques responsables de la fixation des produits des ET sera effectuée à l'aide de méthodes basées sur les profils et les modèles de Markov. Elle met en œuvre des algorithmes permettant des "indels" dans les motifs, car l'extension du motif sur la séquence le contenant peut être légèrement variable.

Détection des régions du génome de D. melanogaster sensibles aux produits de l'élément P

La détection des régions sous le contrôle transcriptionnel des produits de l'élément *P,* reposera sur deux approches génomiques complémentaires : une approche *in vivo* couplée à une approche *in silico*

***Approche* in vivo**

Elle repose sur la comparaison du transcriptome d'une lignée possédant des éléments *P* à celui d'une lignée qui en est dépourvue. Afin de se soustraire à tout effet non contrôlé du contexte génétique, deux contextes génétiques seront utilisés: la lignée ISO1 séquencée par BDGP et la lignée Gruta. La lignée ISO1 a été transformée par nos soins en y injectant des éléments *P* (ISO1-P). La souche Gruta possède déjà une telle lignée transformée: HS2-25. En effet, la lignée Gruta (Argentine) dérive de quelques individus piégés dans une population naturelle en 1950, c'est-à-dire avant l'invasion de l'espèce par les éléments *P*. La lignée HS2-25 a été construite en 1985 par injection d'éléments *P* dans des embryons de la lignée Gruta.

La comparaison des trancriptomes, à l'aide de biopuces, des lignées ISO1 et ISO1-P, ainsi que ceux des lignées Gruta et HS2-25 doit permettre d'identifier des gènes différemment exprimés en présence ou en absence des produits de l'élément *P*. Les comparaisons seront réalisées à différents stades du développement (embryons, larves au stade III et adultes mâles et femelles séparément). Les gènes dont les taux de transcription seront significativement différents entre les deux lignées seront étudiés par analyse moléculaire (Southern blot, PCR, …) pour rechercher d'éventuelles insertions d'éléments *P* dans leur environnement génomique. En effet, certaines différences d'expression pourraient être associées à des insertions de séquence *P* sans pour autant provoquer une modification phénotypique directement identifiable. De plus, les différences d'expression non associées à des insertions d'éléments *P* seront imputables à l'effet de la fixation des produits de *P* sur une séquence présentant un consensus avec le motif canonique. Les régions génomiques correspondantes seront, à leur tour, examinées à l'aide des logiciels d'identification que nous aurons utilisé pour la caractérisation *in silico* des motifs de fixation.

***Approche* in silico**

Elle repose sur la détection dans la séquence génomique des motifs de fixations de la transposase. Les motifs que nous aurons caractérisés seront recherchés directement sur la séquence. L'effet de la présence des produits de l'élément *P* sur la transcription des gènes situés à proximité de ces motifs sera examiné *in vivo*. Ces motifs peuvent aussi être le

résultat de l'insertion d'anciens éléments *P* (préexistant à l'invasion récente à partir de *D. willistoni*). Nous rechercherons donc également les séquences vestiges de ces ET.

Validation des effets des produits de l'élément P sur les motifs détectés

Les interactions séquences-produits de l'élément *P* seront soumises à une analyse génétique. Les régions que nous aurons identifiées comme potentiellement régulatrices seront fusionnées à un gène rapporteur et l'ensemble, inséré dans un vecteur de transformation de la drosophile autre que le vecteur *P* (les vecteurs *piggyBac, hobo* ou *hermes*). L'expression du gène rapporteur sera mesurée en présence et en absence d'éléments *P*. Nous étudierons alors les gènes situés à proximité des séquences ainsi identifiés.

5. BIBLIOGRAPHIE

ANDRIEU O, FISTON AS, ANXOLABÉHÈRE D, QUESNEVILLE H (2004) – Detection of transposable elements by their composition bias – BMC Bioinformatics 5:94

ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ (1990) Basic local alignment search tool. J Mol. Biol. 215: 403-410

ALTSCHUL SF,MADDEN TL, SCAFFER AA, ZHANG J, ZHANG Z, MILLER W, LIPMAN DJ (1997) Gapped BLAST and PSI-BLAST : A new generation of protein database search programs. Nucleic Acids Res 25 : 3389-3402

BAILEY JA, YAVOR AM, MASSA HF, TRASK BJ, EICHLER EE (2001) Recent segmental duplications in the human genome. Genome Res; 11: 1005-1017.

BAO Z, EDDY SR. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12(8):1269-76.

BOURGAIN C, GENIN E, QUESNEVILLE H, CLERGET-DARPOUX F. (2000) Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet. 64 (Pt 3):255-65.

BOURGAIN C, GENIN E, HOLOPAINEN P, MUSTALAHTI K, MAKI M, PARTANEN J, CLERGET-DARPOUX F. (2001a) Use of closely related affected individuals for the genetic study of complex diseases in founder populations. Am J Hum Genet 68(1):154-159.

BOURGAIN C, GENIN E, MARGARITTE-JEANNIN P, CLERGET-DARPOUX F. (2001b) Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. Genet Epidemiol. 21 Suppl 1:S560-4.

BOURGAIN C, GENIN E, OBER C, CLERGET-DARPOUX F. (2002) Missing data in haplotype analysis: a study on the MILC method. Ann Hum Genet. 66(Pt 1):99-108.

CHAO KM, ZHANG J, OSTELL J, AND MILLER W (1995) A local alignment tool for very long DNA sequences. Comput. Appl. Biosci. 11: 147-153

DIZIER MH, QUESNEVILLE H, PRUM B, SELINGER-LENEMAN H, CLERGET-DARPOUX F. (2000) The triangle test statistic (TTS): a test of genetic homogeneity using departure from the triangle constraints in IBD distribution among affected sib-pairs. Ann Hum Genet. 64(Pt 5):433-42.

DIZIER MH, BARBRON MC. (2001a) Triangle test statistic in discordant sib pairs: test of genetic heterogeneity of asthma and atopy in CSGA families. Genet Epidemiol. 21 Suppl 1:S192-7.

DIZIER MH, BESSE-SCHMITTLER C, GUILLOUD-BATAILLE M, SELINGER-LENEMAN H, KAUFFMANN F, CLERGET-DARPOUX F, DEMENAIS F; EPIDEMIOLOGICAL STUDY ON THE GENETICS AND ENVIRONMENT OF ASTHMA GROUP. (2001b) Indication of linkage and genetic heterogeneity of asthma according to age at onset on chromosome 7q in 107 French EGEA families. Eur J Hum Genet. 9(11):867-72.

DIZIER MH, QUESNEVILLE H, BESSE-SCHMITTLER C, GUILLOUD-BATAILLE M, SELINGER-LENEMAN H, CLERGET-DARPOUX F, DEMENAIS F. (2003) Indication of linkage and genetic heterogeneity for asthma and atopy on chromosomes 8p and 12q in 107 French EGEA families. Eur J Hum Genet. 11(8):590-6.

EMANUEL BS, SHAIKH TH (2001) Segmental duplications: an 'expading role ingenomic instability and disease. Nat Rev Genet 2: 791-800.

GUSFIELD D (1997) Algorithms on strings, trees, and sequences. Computer Sciences and Computational Biology. Cambridge University Press. pp. 325-329

GRAY YH. (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. 16(10):461-8. Review.

HOSKINS RA, SMITH CD, CARLSON JW, CARVALHO AB, HALPERN A, KAMINKER JS, KENNEDY C, MUNGALL CJ, SULLIVAN BA, SUTTON GG, YASUHARA JC, WAKIMOTO BT, MYERS EW, CELNIKER SE, RUBIN GM, KARPEN GH (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biology* 2002, 3(12):research0085.1 - 0085.16

HORVATH JE, VIGGIANO L, LOFTUS BJ, ADAMS MD, ARCHIDIACONO N, ROCCHI M, EICHLER EE (2000a) Molecular structure and evolution of an alpha satellite/ non-alpha satellite junction at 16p11. Hum Mol Genet. 9: 113-123.

HORVATH JE, SCHWARTZ S , EICHLER EE (2000b) The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. Genome Res; 10: 839-852.

KAMINKER JS, BERGMAN CM, KRONMILLER B, CARLSON J, SVIRSKAS R, PATEL S, FRISE E, WHEELER DA, LEWIS SE, RUBIN GM, ASHBURNER M, CELNIKER SE. (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 3(12):RESEARCH0084. Epub 2002 Dec 23.

KIDWELL MG, LISCH DR. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution. : 55(1):1-24. Review.

LERAT E, CAPY P, BIEMONT C. (2002) The relative abundance of dinucleotides in transposable elements in five species. Mol Biol Evol. 19(6):964-7.

LEWIS SE, SEARLE SMJ, HARRIS N, GIBSON M, IYER V, RICHTER J, WIEL C, BAYRAKTAROGLU L, E BIRNEY E, CROSBY MA, KAMINKER JS,

MATTHEWS BB, PROCHNIK SE , SMITH CD, TUPY JL, RUBIN GM , MISRA S, MUNGALL CJ , CLAMP ME (2002) Apollo: a sequence annotation editor. Genome Biology 3: research0082.1-0082.14

MCCARTHY EM, MCDONALD JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19(3):362-7.

NASSIF N, PENNEY J, PAL S, ENGELS WR, GLOOR GB (1994) Efficient copying of nonhomologous sequences from ectopic sites via *P*-element-induced gap repair. Mol Cell Biol 14:1613-1625.

NOUAUD D, ANXOLABÉHÈRE (1997) *P* element domestication: a stationary truncated *P* element may encode a 66-kDa repressor-like protein in the *Drosophila montium* species subgroup. Mol. Biol. Evol. 14: 1132-1144

NOUAUD D, BOEDA B, LEVY L, AND ANXOLABEHERE (1999) A *P* element has induced intron formation in *Drosophila.* Mol. Biol. Evol. 16: 1503-1510

NOUAUD D, QUESNEVILLE H, AND ANXOLABÉHÈRE D (2003) Recurrent Exon Shuffling Between Distant *P*-Element Families. Mol. Biol. Evol. 20(2) : 190-199

QUESNEVILLE H, ANXOLABÉHÈRE D (1997a) GENOOM: a simulation package for GENetic Object-Oriented Modeling. Annals of Human Genetics 61, 543.

QUESNEVILLE H, ANXOLABÉHÈRE D (1997b) Simulation of P element horizontal in Drosophila. Genetica 100, 295-307

QUESNEVILLE H, ANXOLABÉHÈRE D (1998) Dynamics of transposable elements in metapopulations: a P element invasion model. Theor Pop Biol 54, 175-193

QUESNEVILLE H, ANXOLABÉHÈRE D (2001) Genetic Algorithm-based Model of Evolutionary Dynamics of Class II Transposable Elements - J. Theor. Biol, 213, 21-30

QUESNEVILLE H, DIZIER MH, CLERGET-DARPOUX F. (1999) Departure from the triangle constraints in discordant sib pairs: a test for genetic heterogeneity. Genet Epidemiol 17 Suppl 1:S685-9.

QUESNEVILLE H, NOUAUD D, ANXOLABEHERE D (2003) Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genome. J Mol. Evol. 57 : 1-10

SAMONTE RV, EICHLER EE (2002) Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3: 65-72.

SHIELDS DC, SHARP PM. (1989) Evidence that mutation patterns vary among Drosophila transposable elements. Mol Biol. 20;207(4):843-6.

6. ANNEXES