UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

GENE ANNOTATION IN *HELICONIUS NUMATA*
COLOURING THE *BLACK BOX*

Simone Fernandes Delgado

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

2010

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

GENE ANNOTATION IN *HELICONIUS NUMATA* COLOURING THE *BLACK BOX*

Simone Fernandes Delgado

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

Dissertação de mestrado orientada pelos professores doutores Sara Magalhães e Mathieu Joron

2010

Index

Resumo *1*

Abstract: *1*

Introduction *3*

Material and Methods: *10*

BAC clones 10

Annotation 10

ESTs 10

Blast search 11

Archiving the models 13

Primers 14

Practical Approach 14

PCR 16

RNA extraction 16

Results *18*

Models Tested 23

Conclusion and Discussion: *26*

Gene 41 27

Gene 53 27

Gene 54 27

Gene 23 28

Gene 24 28

Gene 25 29

Gene39 29

Gene 40 29

Perspectives *32*

References *33*

Appendix *36*

Resumo

O presente projecto teve por objectivo a identificação e anotação de genes que se sabe estarem ligados num arranjo de genes inquebrável por recombinação e que controla a padronização das asas da borboleta *Heliconius numata.* Este complexo é homologo posicional de complexos em outras espécies de *Heliconius*, nomeadamente o complexo Yb/Sb em *H. melpomene* e *Cr* em *H. erato.* Este "supergene" é visto como um *hotspot* de desenvolvimento *(sensu* Richardson & Brakefield 2003*)* que controla divergência fenotípica entre espécies próximas e convergência fenotípica entre espécies mais distantes. Em *H. numata* o supergene sofreu rearranjos genómicos que se pensa estarem correlacionados com a preservação das combinações alélicas necessárias para manter os fenótipos discretos característicos das relações miméticas locais (*mimicry rings*) que estes organismos estabelecem na natureza. Estes rearranjos podem estar a modificar ou perturbar expressão genica e é, portanto, de extrema importância identificar tais genes bem como compreender a sua estrutura, nomeadamente no que diz respeito a eventuais eventos de clivagem alternativa que possam estar correlacionados com as diferentes raças locais observáveis na natureza. Usando uma combinação de ferramentas bioinformáticas e sequências de transcriptoma de raças com fenótipos diferentes, modelos de genes serão estabelecidos para desenhar primers específicos de modo a amplificar os genes modelizados e testar hipóteses relativamente a diferenças na sua estrutura e distribuição na região de interesse.

Palavras-chave: Sintenia, Clivagem Alternativa, Supergenes/ Genes switch

Abstract:

The present project aims at the identification and annotation of the specific genes known to be in a cluster of tighly linked genes known to control wing patterning in *H. numata*. This cluster is positionally homologous with cluster in other species of *Heliconius*, namely Yb/Sb in *H. melpomene* and *Cr* in *H. erato.* This "supergene" is seen as a *developmental hotspot (sensu* Richardson and Brakefield 2003*)* controlling both phenotypic divergence between closely related species and convergence between more distantly related species. In *H.* numata the supergene suffered genomic rearrangements that are thought to be preserving the necessary combination of loci to attain the concrete phenotypes typical in mimicry relations these butterflies establish in nature. These rearrangements may be disturbing or modifying gene expression. Therefore, it is extremely valuable to identify the genes in the genomic region of interest, and address eventual alternative splicing events that could be associated with the different locally adapted races. Using a combination of bioinformatics' tools and transcriptome sequences from two phenotypically different races, I generated gene models in order to design primers and amplify these genes, testing hypothesis regarding splice variants and validating the models.

Key words: Sinteny, Alternative Splicing, supergenes/switch genes

Acknowledgements

I would like to thank Mathieu Joron for receiving me in his work team and and for the trust and help with a task that will be necessary and important for further future work. Thank you for the motivational talks, the constructive critics, the encouragement during the process of this project and regarding future work in my scientific career.

I would also like to give a special thanks to Lise Frezal for the valuable help regarding practical work and specially for the support and the amazing comradeship.

Yann Le Poul, thank you for you support and friendship. Many thanks to the rest of the team for the support and for setting such a pleasurous mood to work in.

Gostaria de agradecer igualmente aos coordenadores do Mestrado de Biologia Evolutiva e do Desenvolvimento, a professora Maria Collares Pereira e o professor Élio Sucena bem como à minha orientadora interna professora Sara Magalhães.

Agradeço à minha família e amigos em particular aos meus tios Donzília e Daniel que me acolheram e me possibilitaram cumprir o sonho de viver em Paris, à minha mãe pelo apoio incondicional e inigualável de sempre que foi tão importante nesta etapa fora de casa.

Introduction

Evolution of physical characteristics happens through genetic change over successive generations. Although is not always clear which genes are involved, the ones in embryonic or larval development are likely candidates since they control development of animal's characteristic form and features (36).

Traditionally it was believed micromutations were driving Evolution, but the similarity between some dramatic mutations and evolutionary transitions supported a role for macromutationism (16). Arguments on both sides have been biased by misconceptions of the developmental effects of such mutations. Mutation of key developmental genes was thought to imply large pleiotropic effects. (36) We now know that some mutations such as those in *cis*-regulatory regions of genes have few or no pleiotropic effects constituting a "safe" source of morphological evolution (36). This new perspective assists in identifying of specific mutations and in determining how they alter gene function and generate phenotypic variation. These data can provide the missing link between molecular and phenotypic variation in natural populations, the so called *black box*.

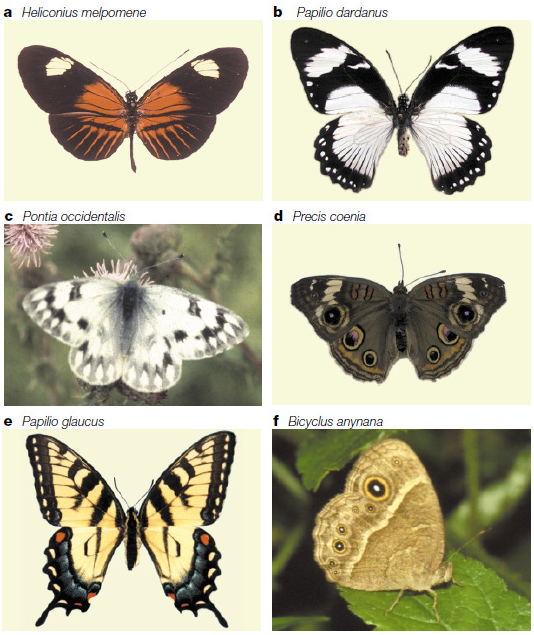
Most work has focused on major events in body plan diversification with fewer studies of intraspecific or interspecifies differences (4, 35). It therefore remains unclear if there are distinct levels of diversification, or if the higher levels can be extrapolated from microevolution.

Rather than accumulation of entirely new mutations, because of the time it would take, it seems more likely that the evolution of novelty emerges from modification in already existing sequences, i. e. tinkering of the available genomic toolkit (36) Sucena *et al.* 2003 demonstrated that, indeed modification in few developmental key elements can drive parallel evolutionary changes independently in different species, meaning that existing diversity in these sequences allows for finding similar solutions for the same problems, a solution more rapidly attained by mutations of major effect.

This raises the question: Does adaptive divergence typically proceed by the substitution of many genes of minor effect (17) or can genes of large effect contribute to adaptation and speciation (13) Molecular genetic studies are increasingly addressing the question about the relative contribution of gradual evolution driven by natural selection versus macromutationism.

The field of evo-devo aims for a mechanistic understanding of origins and evolution of diversity: how are developmental pathways modified to produce novelty, whether there's a tendency in which gene classes usually underlie adaptive change and if there's a general trend in their modes of action. For instance, homeotic genes have provided us insights into major transitions in animal body plans and how genetic networks are modified in diverse taxa (10) and are a good example of genetic pathways conserved/shared in organisms with very different body plans (1). This gives provides a common ground in which to understand diversity in a broad scale, but to address adaptive change and understand how is driven by natural selection, it is necessary to compare closely related species in which the diversification is recent. (2, 3)

It has recently become apparent that speciation is often caused by normal processes of adaptive differentiation under divergent natural selection (19) so that macroevolution might be simply extrapolated from microevolution.



1 - Different wing patterns in butterflies

Butterflies' colour patterns (Fig. 1) provide a good model system given their amazing natural diversity, their recent adaptive radiation and a long history of studies from the molecular details of insect colorvision (7) to analysis of human impact on biodiversity (3). Also, this organism provides us with the phylogenetic and genetic level of variation relevant to advance our understanding of morphological trait evolution.

From an evolutionary perspective, butterfly colourful wing patterns are interesting because their adaptive value is known in most cases, and they are visual products of (sometimes strong) natural selection. Wing patterns differ between and within species (genetic polymorphisms, polyphenisms (21, 24), sexual dimorphism, etc.). Ecologically, their function is intra and/or interspecific communication, including aposematic warning coloration (26) to avoid predation (25) as well as recognition, mate localization and sexual selection (19).

Patterns of phenotypic and genotypic variation in *Bicyclus'* typical eyespots have been explored within and across species (28) and several genetic pathways known from *D. melanogaster* wing development have been implicated in the formation of eyespot on *Bicyclus* wings (5). Lepidoptera's

development has been shown to involve a number of genetic pathways known other insects namely formation of particular wing pattern elements relying on pathways involved in overall wing formation in *Drosophila* (5) and wing scale production recruiting a pathway from *Drosophila* bristle development (18)
In order to understand butterfly wing pattern patterning, we need to identify which genes are responsible for such patterns and which pathways are they involved in.

Undoubtly, knowledge from *D. melanogaster* wing development studies is helpful but Lepidoptera and Diptera are diverged enough and their development is quite different (*e.g.*, dipterans have a single pair of wings and colored scales are a morphological innovation of Lepidopterans) and therefore is unlikely that all the genes involved in butterfly wing pattern formation will be genes known from *Drosophila* wing development (31)

This suggests that different pathways or new genes in the same pathways should be implicated in the development of butterfly wing patterns. More studies of the developmental genetics of pattern formation in different species combinated with gene-mapping approaches will help answering to the major questions: What genes are involved in the process of patterning? What are the developmental pathways in which these genes take part?

Heliconius butterflies (Nymphalidae: Heliconiinae) include several examples of strong divergence between geographic races of the same species and near-perfect local mimetic convergence between species, providing a link between molecular genetics and adaptive evolution (Fig. 2).

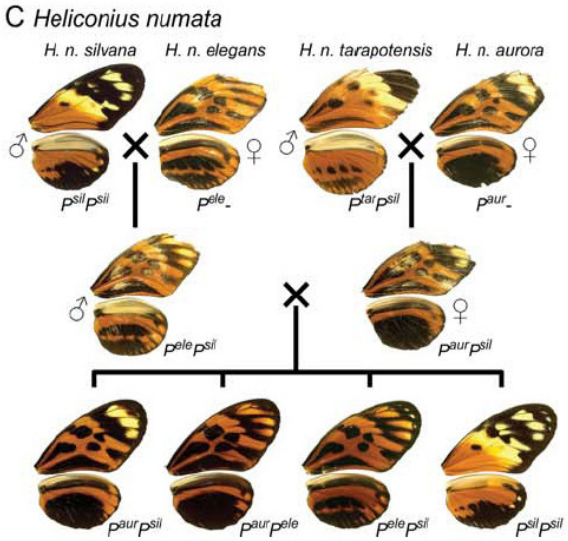
This group presents extensive natural variation in colour pattern and a long history of ecological and evolutionary research (25). Studying colour polymorphism in these species allows to focus on phenotypic variants that are influenced by known modes of natural selection (19, 22, 25). This vivid colour patterns warn predators of their unpalability and are presumably related to their evolutionary history in association with cyanogenic foodplants in the Passifloraceae (21). Also, radiation in Heliconius colour patterns couples divergent evolution and multiple independent cases of convergence in müllerian mimicry rings, representing different evolutionary timescales.

1 - Phenotypic divergence between closely related H. numata and H. melpomene contrasts with phenotypic convergence between more distantly related H. melpomene and H. erato

The link between phenotypic shifts in pattern and macroevolutionary diversification make the radiation in the mimetic wing patterns of Heliconius butterflies a perfect model system to explore the nature of constraints, bias, optimality, and chance in morphological change (20).

Homologous genetic pathways (29) or a limited number of loci controlling colour pattern shifts (24) seem to have a general role in convergent mimicry (20). By crossing individuals of the same species displaying different wing patterns, dominant and recessive traits can be observed segregating in broods (Fig.3). Mendelian crosses 58 showed that inheritance in natural populations of *Heliconius* (as well as in *B. betularia*, *P. dardanus*) involves a small number of genes of major effect (11).

Nevertheless this low mapping resolution doesn't allow to distinguish between a single locus and several tightly linked loci.



3 - Controlled crosses in *H. numata*

The repeated involvement of homologous loci in evolution of convergent phenotypes could suggest

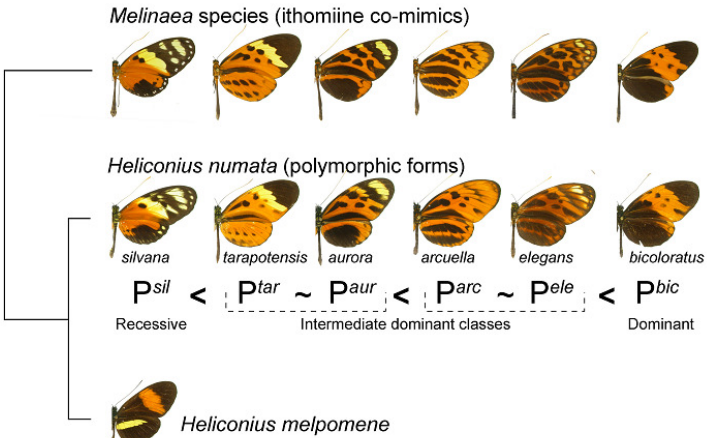
strong developmental constraints on adaptive evolution (34) but, the same loci are also recruited in divergent evolution which means they may be important not only in convergence but in phenotypic evolution in general (21). It seems intuitive to predict these will correspond to developmental genes acting upstream scale maturation pathways, regulating spatial expression and controlling development of morphology and pigmentation in adult wing patterns (29).

High-resolution linkage maps and molecular markers transferable between species allow to test hypotheses about genetic homology between mimetic species and further explore the architecture underlying convergent and divergent evolution in *Heliconius.*(20)

Work on early stages of pattern formation in Heliconius (32) hasn't clarified any association between gene expression patterns and wing pattern elements seen in the eyespots of other butterflies (3) Part of the difference in patterning between butterfly groups may be due to temporal changes in conserved pattern-formation processes. As an example, the *Notch/Distal-less* (N/Dll) pathway is associated with intervein elements and eyespots in Bicyclus, and is truncated in species lacking eyespots including *Heliconius* (32)suggesting that pattern variation in *Heliconiines* can either be associated with regulation of earlier stages of the N/Dll process, or involve completely distinct pathways. The recent discovery of a veinless mutant further supports this: *Heliconius* patterns develop independently of wing (20) disproving Nijhout's (1991) hypothesis of a common 'nymphalid ground plan' and implying *Heliconius* patterns are probably result from a distinct, unexplored, patterning system.

With the exception of tight linkage between *wingless* and the white/yellow colour switch *locus K* in *H. cydno* (20), several loci important in *Drosophila* wing development (*apterous, wingless*), in Bicyclus eyespot specification or in scale pigment synthesis (*vermilion, cinnabar*), don't seem to be linked to pattern switch genes in one or more *Heliconius* species (22). This result implies that novel or unexpected genes or pathways could be involved in pattern specification for these species.

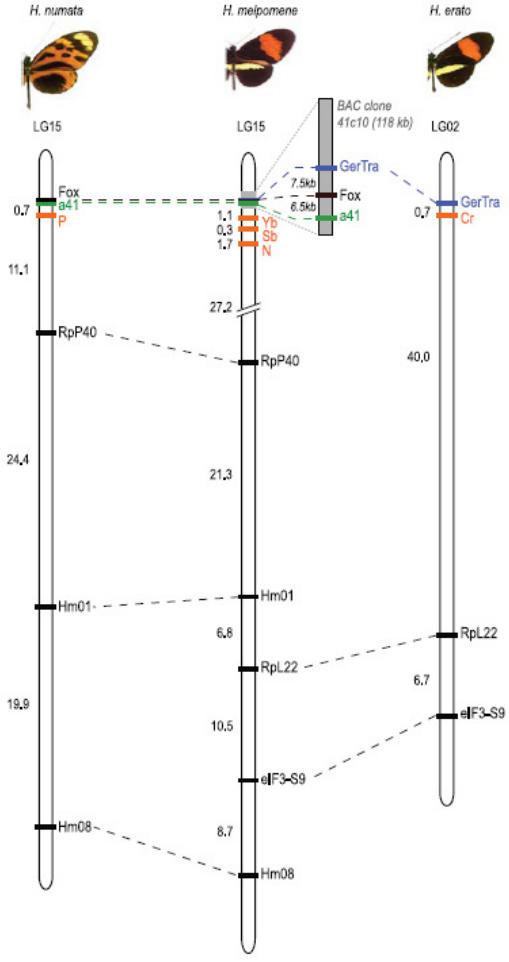
Mimicry polymorphism in *H. numata* patterns is inherited entirely at a single Mendelian locus, P. Populations are locally polymorphic, and nine distinctive alleles have been identified for the P locus (8).These present a clear hierarchy of dominance relationships (Fig 4.) (21), as might be expected in order to prevent the segregation of intermediate nonmimetic phenotypes in wild populations. But, because, occasional recombinant phenotypes do occur, P locus is clearly a tight cluster of genes, or ''supergene''. (8)



2 - Relationship of dominance in *H. numata* polymorphic races

Homology of these clusters implies, rather than a constraint, an extraordinary "jack-of-all-trades" flexibility of homologous colour pattern loci in closely related species. One of the elements that controls geographic variation in *H. melpomene* and *H. erato* has somehow taken over control of the entire pattern in H. numata. Although evolution of linkage between unlinked *loci* is not likely, a gradual reduction in recombination between already linked elements seems plausible presumably facilitated by the fact that these regions already have major phenotypic effects on different parts of the wing in the ancestral species. The linked elements Yb, Sb and N in H. melpomene might have been brought closer together to avoid unfit intermediate genotypes in polymorphic H. numata populations. This complex is interpreted as a 'developmental hotspot' (33), playing a disproportionate role in response to different selective pressures in both divergent and convergent adaptive evolution. This challenges the idea that shared genetic architecture (34) constrains morphological diversification (14). Instead, natural selection has shaped

In contrast, geographic variation in H. melpomene is inherited at multiple loci on 4 chromosomes. Comparative linkage mapping in H. numata and H. melpomene has shown that P is a positional homologue of the cluster of loci Hm Yb/Sb/N on chromosome 15 in H. melpomene and the locus Cr in H. erato (Fig. 5).



3 - Positional homology of loci controling wing patterns

a developmental switch mechanism that responds to different mimetic pressures, producing locally adapted highly divergent colour patterns. This switch might represent a number of *cis* regulatory elements of a single gene a cluster of duplicated genes with divergent (9), or a cluster of nonparalogous but functionally related genes (20)

Recently, local chromosomal rearrangements have been discovered in *H. numata* in association with the P locus (Joron, personal communication). These rearrangements disturb recombination and lock together the variation at neighbouring loci, providing a mechanistic hypothesis for the formation of the "supergene". However, this may also disturb and/or modify the expression and splicing of the genes near the breakpoint

A detailed analysis of the structure of the genes affected by the rearrangements at the P locus is necessary to derive hypothesis regarding their differential expression in different gene arrangements.
Here, I report the detailed study of the subset of genes most likely to be affected by rearrangements and be involved in the development of alternative mimetic morphologies on the adult wing.

Material and Methods:

From previous work, recombinational mapping allowed to identify the genomic region associated with pattern variation in *H. numata*. Although this genome is not available, a genomic library of Bacterial Artificial Chromosomes was built and screened to identify and sequence the genomic region of the P locus. BAC sequences revealed different gene arrangements. The annotation of those BAC sequences is now needed to study gene structure and variation, especially around rearrangement breakpoints.

BAC clones

These BAC libraries were built using 5 individual larvae from a polymorphic population in NE Peru, meaning each piece of DNA in the library is a piece of one of 10 different homologous chromosomes. The average insert size is 115 kilobases, and there are 18432 clones, so knowing that the haploid genome is estimated to be 319-megabase long, the estimated coverage is 6.6X (i. e. on average each unique portion of the haploid genome is represented on 6.6 different clones). The genomic region of interest, *locus P,* spans over at least 5 of these BAC clones, some of them overlapping and contains a set of predicted genes based on previous work (15).

Annotation

I used these BAC libraries for *H. numata* as a starting point to annotate and describe the genes of interest, i. e. to locate the genes and ORFs (open reading frame), identify their putative orthology and function and describe their (intron-exon) structure based on some transcriptome sequences.

ESTs

I used Expressed Sequence Tag sequences from phenotypical different *H. numata's* populations:

Tarapoto including tarapotensis and bicoloratus individuals ($P^{tar} < P^{bic}$)

Yurimaguas population: mostly aurora and isabellinus individuals ($P^{isa} < P^{aur}$).

These EST were made from multiple individual wing discs, at the time the pattern is believed to be specified between 5th-instar larvae and late pupae, allowing, therefore to address gene expression during this period. Since they came from different populations, this gives information on whether there are differences in gene expression and structure between both populations and if there are, our goal is to determine if these correlate to wing pattern specification. They were sequenced using 454-

pyrosequencing technology giving about 300,000 reads on average 200bp long per library, i. e. a total of 600,000 reads. Because of their short length, these were assembled in contigs to identify gene objects in the transcriptome and annotate them by similarity.

The clustered ESTs are useful to locate gene regions and predict orthologs but they can't be used for detailed annotation of individual gene models. For this we need to use the raw reads, each one being a sequence from a single transcript. Raw reads also contain polyadenilation tails (the ending portion of unprocessed mRNA, which are trimmed for the clustering in gene objects) and are therefore informative about transcript termination, which may be variable and must thus be recorded to generate accurate gene models.
Both EST assemblies and raw reads were used to describe the gene models for the regions of interest. For this I used bioinformatics tools to compare genomic sequences (BAC) for the region of interest and the large curated databases made from the ESTs as well as DNA and protein sequence repositories such as GenBank, Uniprot and model organism genomic projects (*e. g. Bombyx mori*, the model Lepidoptera).

After identifying the genes in the BAC sequence, I used a set of bioinformatics annotation tools as well as annotated databases to gather as much information as possible about these genes and design the corresponding models.

One of the most valuable tools in Biology is Basic Local Alignment Search Tool, which performs searches of similarity between a query sequence (either proteic or nucleic) and every sequence in a database using the algorithm of Altschul et al. (1990). The level of similarity is evaluated by the probability of similarity due to chance at the sequence level which is usually exposed as an exponential value (e-value).

Blast search

The first BLAST was performed using NCBI (http://blast.ncbi.nlm.nih.gov/) tools with command lines, to compare the EST sequences with the BAC genomic sequences. To this end, I created a .bat file that contained the command lines and that run all the searches in the different databases at once (see appendix). The output of this search gives the information about which ESTs correspond to each gene.

Using blastall program, we can specify the parameters of the search; here I used mainly tblastx that compares the six-frame translations of a nucleotide query sequence, (i. e. the BAC sequence) against the six-frame translations of a nucleotide sequence database (ESTs). I blasted the genomic sequence of a gene region onto the EST database to recover the reads with a high similarity to the query sequence and that therefore correspond to the gene of interest. The recovered reads which e-value is above ten (e^{-10}) were then converted to fasta format and aligned to the genomic sequence in CodonCode software. An easy to use manual of this program can be found at: http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/blastall_all.html

Because the transcriptome sequencing and our knowledge of genome biology for this non model is only partial it is important to compile as much information as possible (similarity of sequences, putative domains, putative function based on sequence similarity or experimental studies in homologues, etc.). There are several automated annotation tools that compile information from databases in different groups of organisms and generate gene predictions based on motifs most usually associated with splicing sequences, methionine codons, stop codons, low complexity sequences, etc.. Since I'm working with a non-model organism I needed to avoid bias that could arise from the use of different sources of information and therefore defined a hierarchical ranking (Table 1) of the sources, generating a canonical model in which to test the instances in which this information showed discrepancies.

1 Different sequence data sources

Order of importance ↑

Data	**Advantages**	**Disadvantages**
Blastx	Full length genes	Not specific
EST contigs	Find new genes; *H. numata* specific	Partial sequences
EST raw reads	*H. numata* specific splicing information	Short length
Non insects	More data	Data from diverged organisms
Ab inition predictions	Find new genes	Bias

Nevertheless, these automated tools give a good idea of where and how long a gene might be, complementing the raw information given by the ESTs.

The *Heliconius* BAC sequences for the *P* locus were individually annotated with the clustered EST

contigs and masked using the ReRep repeat database (a database of repetitive motifs), using the annotation pipeline MAKER (6). This program identifies repeats, aligns ESTs and proteins to a genomic DNA sequence, producing *ab initio* gene predictions, and automatically synthesizes the data into gene annotations with evidence-based quality indices. Therefore, it provides an effective means to annotate an emerging genome and to create a genome database. It can be used to annotate individual contigs and BACs. The MAKER pipeline resulting files can be analysed and viewed with the Apollo Genome Annotation Curation Tool version 1.9.6

Gene predictors used here were:

- In the Maker pipeline: Semi-HMM-based Nucleic Acid Parser (SNAP) using a Hidden Markov Model optimized to the *Bombyx mori* genome.

- The automated *Bombyx* gene prediction tool 'Kaikogaas' (http://kaikogaas.dna.affrc.go.jp/), a gene annotation web system based on parameters determined from the *B. mori* genome.

Archiving the models

I used Artemis software to create features (annotated tracts of DNA, for instance exons, introns, 5'UTR, etc…). Here I used the .fas BAC sequences and looked manually where each automated prediction (both Maker and Kaikogaas) would align as well as EST assembled contigs and raw reads from *H. numata* and *H. melpomene.* Using a code of colours it was possible to concatenate organized information from all the sources and determine which features where robustly supported and which should be tested

Gene Models

A schematic representation of the information regarding each putative gene, was generated in a graphic software to visualize more clearly the differences between predictors as well as differences in the predictions derived from the genomic sequence of different clones (for the cases where genes are represented in more than one).

Because the locus encompasses more than 30 genes, the present project was targeted to study in detail the gene models for the regions surrounding the rearrangement breakpoints. These genes are numbered in agreement with previous work in *H. melpomene* (15) and named according to the best

hit in *Drosophila'*: gene 21, 22, 23/ATP-binding protein, 24/LRR, 39/Licorne, 40/ERCC6, 41/penguin, 53/l(2)gl and 54/zinc finger protein.

Once the genes were identified in the genomic sequence, the corresponding ESTs were aligned and models have been designed the resulting putative models were available for testing, i.e. validation especially it the alignments of different raw reads conflicted with each other or with known genes and proteins. Using wing tissue from larvae of known wing patterns phenotype to perform PCR (Polymerase Chain Reaction) to amplify the genes from the retro transcribed transcriptome (cDNA). To address splicing variation I therefore generated cDNA from individuals of both populations to amplify the genes, and compare the lengths expected based on the models.

Primers

To generate the primers to amplify the genes and examine their structure, I identified the problematic regions, i.e. the regions where the ESTs don't support the predictions or regions in which predictions don't match. The primers were then designed to span the exons and/or introns in which such questions arose. In designing the primers, it is important to take into account that the primers won't hybridize with each other, that they're not too long (because that would increase their specificity) and that they align in regions well supported by the data. Taking the coding sequence and choosing the target portion of the sequence, the software Primer3 http://frodo.wi.mit.edu/primer3/ predicts the best pair of primers.

The length of the fragments to be amplifying using the chosen primers was calculated by aligning them against the BAC clones and canonical models. The primers were ordered in www.eurofinsdna.com selecting unmodified DNA oligos, salt free, lyophilized and concentrated at 0,01 μmol.

Practical Approach

In order to make sure the primers designed were indeed good to amplify the genes of interest in cDNA, they were first individually tested in genomic DNA from three individuals mj02-731(bicoloratus), mj02 – 268 (tarapotensis), mj02-163 (tarapotensis) in a trial PCR.

To determine if the different phenotypes are due to differential expression of the genes in the *locus P*, and considering the models designed and supported (or not) by the EST sequences, we decided to extract RNA using commercially available kit from Qiagen to obtain from it the corresponding cDNA. To check for genomic DNA contamination we did electrophoresis with genomic *versus*

cDNA and checked for different length of the amplified fragments.

PCR

PCRs were performed in the thermocycler machines from Gene Amp ®, PCR system 2700 and, G-storm (which allows for setting different temperatures for the same cycle)

PCR was performed using standard conditions:

- 94º C (2 to 4 minutes)
- 94º C Denaturation (20 seconds)
- 35 x Tm: [(4(C+G)+2(A+T)) – aprox 4º C]. Tm is usually estimated by the commercial distributor (Eurofins in this case)
- 72º C Elongation (1min/kb)
- 72º C (10min)

end

- 4º C or 10º to 15º C if the PCR products have to wait longer (or even overnight) in the machine.

Conditions of Tm (melting temperature) and elongation time were optimized for each pair of primers (table in appendix), lowering when necessary to a minimum of 51ºC, and elongation time 1min per Kb of the expected fragment.
Since the primers were designed in the coding sequence, and in some cases the introns are extremely long, which makes the amplification difficult, I used primers available in the teams' stock known to align well (Hn21_R1 was tested with ex6-F985 and Hn41_pen_F1 was tested with HnPenF824).
The amplification reactions were performed using 10 µL as a final volume with 0,6 µL of DNA and 1 of cDNA.

For the electrophoresis, agarose gel concentration was 1.2% (1.2 g/100mL TBE 1X) and charged with 5 µL (3 µL PCR product + 2 µL of bromophenol blue) and 3microL of Ladder 100bp.

RNA extraction

Extraction was performed using the Quiagen RNAeasy kit. To test and adapt the protocol, a first extraction was performed in wing disc of pupae from two individuals *tarapotensis*: mj07-2145 (posterior wing) and mj07-2156 (anterior wing). This test extraction benefited from the fact that the

wing discs are much more developed in this stage making the extraction process easier.
Using anterior and posterior wing discs from 5 individuals at the L5 stage (mj07-2148 and mj07-2149, predicted to be from morph silvana (sil x sil), mj07-2152 – A86 bic female (from arc mother) x wild bic male, and mj07-2153 and mj07-2154 from Ta (C2)), we performed the RNA extraction to obtain the corresponding cDNA.

The samples were preserved in a stabilizing solution and were then transferred to RNAse free 1,5 mL eppendrofs to start the extraction by crushing the biological material with the help of the buffer (from the kit) in dry ice.

In order to have good control for the length of the fragments amplified, we also performed DNA extraction from the same individuals in both developmental stages. This extraction was performed using a commercially available kit from Quiagen (details in appendix)

Quantification of the nucleic acids is important in order to calculate the precise proportions of reagents during the course of purification and the RT-PCR. This quantification of the nucleic acids was performed using a nanophotometer from Implen. After each extraction and labelling of the cDNA and DNA, the samples can be stored at -20º.

Results

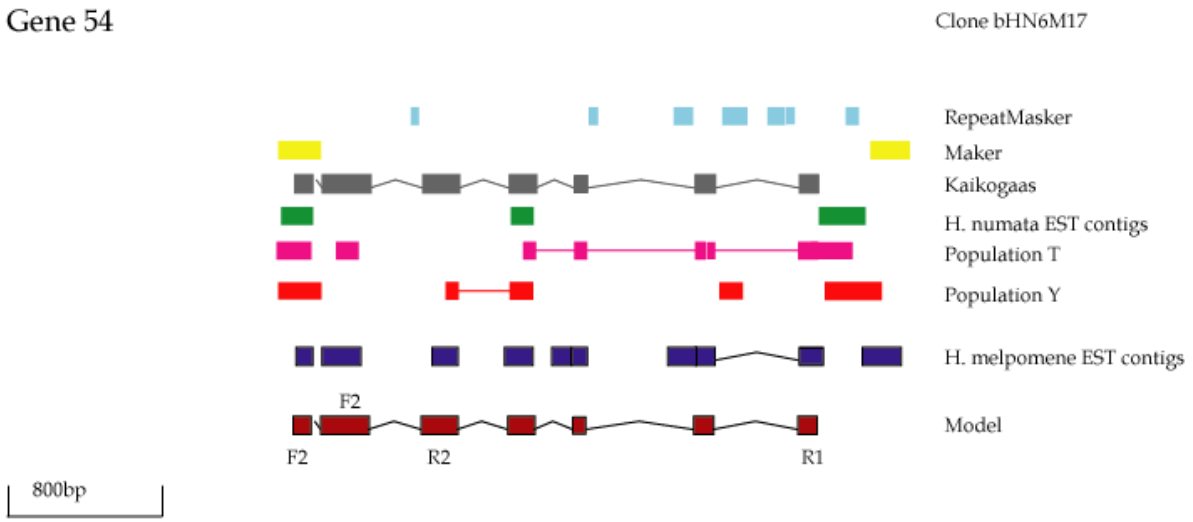
Here I present the results of the manual annotation of the genes of interest. Starting with automated predictions for each gene, I performed alignments of ESTs from two populations of *H. numata* and *H. melpomene* against the genomic sequence of *H. numata.* The ESTs used included assembled contigs as well as raw reads, which allowed to address eventual events of alternative splicing between populations that can possibly be associated with the observed differences in wing patterns.

Also, this observation of sometimes conflicting information generated by the automated pipelines *versus* the expression data allows to address the true length of genes and the exon/intron barriers, something that can be neglected in the automated generated models. This way, it was possible to correct or confirm each model in a critic manner, keeping in mind the hierarchy of the available data sources.

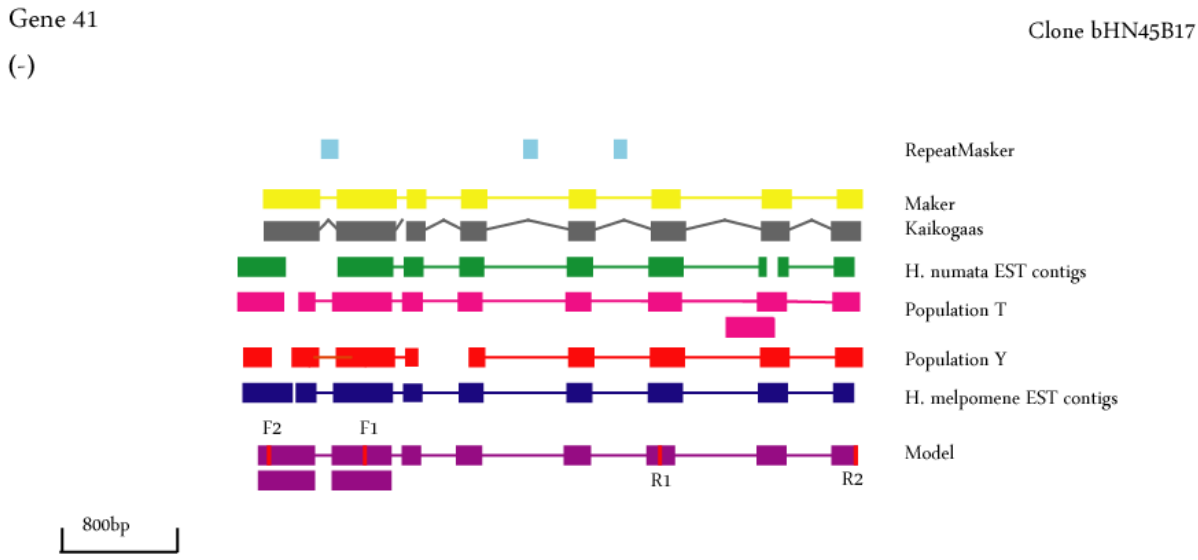
The importance of this analysis comes from the fact that the databases are usually biased towards the most commonly used organisms in the field and even for the trained pipelines, they are programmed using parameters optimized for the moth *Bombyx mori* and, despite the fact both are Lepidopterans, these conditions might not be optimal for butterfly genome analysis.

By comparing the automated predictors it is notable that sometimes they don't match (between Kaikogaas and Maker and even in Maker depending on the blast used). Complementing this with the ESTs for which the alignment is good is not only important, but necessary to determine the robustness of such predictions. Also, these are pipelined data, meaning they neglect a good part of the existent variation that can be observed upon manual alignment of the mentioned ESTs.

Just by looking carefully for each putative gene's well aligned raw reads, I found several instances where they seem to suggest alternative splicing events, i.e. sequences with poly A tails aligning in a regions where other sequences don’t have a poly A tail, for instance. This supports the notion that indeed alternative splicing events happen in the region and considering the rearrangements on the locus P, it becomes even more important to identify where these events take place and investigate their effects.



When creating the gene models, I used different code colours facilitate the finding of differences in support from the two population's ESTs: green label for the numata EST (general mixed pool), pink for ESTs from the assembled contigs and raw reads of T population, red for the assembled contigs and raw reads from population Y and purple for *H. melpomene* assembled contigs and raw reads. This allows to check if there is a striking difference between populations in terms of support of the features. The yellow features depict the maker prediction for the putative exons and the grey represent the autopredgenes from Kaikogaas. Both of these predictors are based of a battery of data (snap, blastn, blastx, est2genome, etc.) and trained in *Bombyx* sequences.



4 - Gene Model for gene 41

Here I present one single example, the gene 41, but all the remaining gene models generated are represented in appendix. In this example we can see clearly that the features predicted by both

Kaikogaas and Maker are strongly supported by the EST data. All the features were blasted against NCBI databases in order to confirm their putative identity in each gene. For cases where there are poly A tails before the putative end of the gene, as the example above, these were properly noted and primers were designed to include such features in order to test for splicing events.
The resulting gene models hint to which features should be tested in PCR in order to confirm splice variants. From bioinformatics' analyses, I generated the following hypothesis to be tested:

Gene	Conflict in Predictions	Support from ESTs	Alternative Splicing	Putative Role
21	+	weak	-	Unkown
22	0	strong	-	Enoyl-Coa Hydratase
23	+	strong	in1, ex5, ex7	ATP-binding protein
24	+	strong	in1, ex2	LRR
25	++	weak	ex3, ex5, ex10	Fizzy protein
39	0	strong	-	Licorne
40	+	weak	in1, ex3, ex4	ERCC6
41	0	strong	ex2, in3	Penguin
53	++	weak	No data	Lethal (2) giant larvae
54	+	weak	in1, in2, ex5	Zinc finger protein

The primers were individually tested in genomic DNA before the RNA extraction reaction to ensure that the sequence, the features to test and the cycles of PCR were good. In this phase of the work it was already noticeable that some of the fragments were too long to amplify from genomic DNA, for example for gene.

The primers and gene models tested are in the following table:

Gene	F- Primer	R-Primer	Model (bp) genomic	cDNA	MEGA Alng 11J7mel		To Test
21	Hn21_HP_F1	Hn21_HP_R1	2550	1419	2762		ex4 and ex7
					38G4	**45B17**	
	F85	R913	1291	829	neg	1285	ex4 and ex5 (end of the gene)
23	F223	R464	369	242	367	364	ex2 and ex3
	F223	R913	1065	691	1061	1060	ex4
	F521	R1217	1466	697	1482	2553	putative in1, ex2
24	F711	R1217	1276	507	1292	2363	ex2
					7C9		
39	F291	R873	2077	583	2094		ex3, ex5, ex11
40	F689	R1272	969	584	968		put.in1 and ex2

The remaining combinations of primers and hypothesis to be tested in future studies are in appendix. Due to time limitation, not all the models and hypothesis were tested.

For those that were, the results are represented in the following table and gel pictures in appendix and the PCR results were confronted with the gene models for each case.

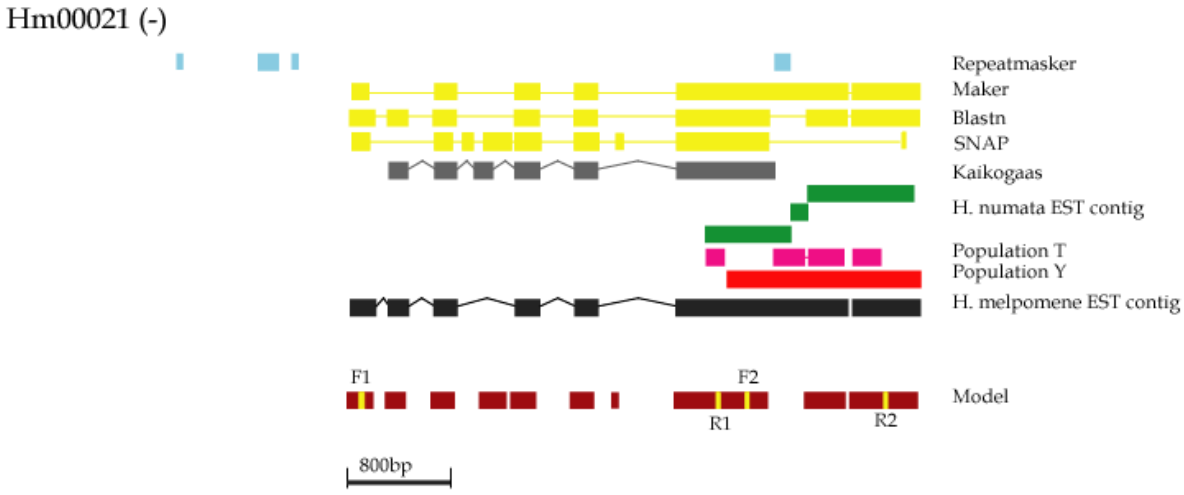
		predicted length (pb)		H. *n. bicoloratus* (dominant form)						H. *n. silvana* (recessive form)				H. *n. tarapotensis*				
				mj07-2145		mj07-2152		mj07-2156		mj07-2148		mj07-2149		mj07-2153		mj07-2154		
gene	primers	gDNA(g)	cDNA('c)	g*	c*	g*	c*	g*	c*	g*	c*	g*	c*	g*	c*	g*	c*	interpretation
LRR	F711/R1217	1300	na	1300	500	1300	500	1300	500	abs	500	1300	500	1300	500	1300	500	no alternative splicing
	F521/R1217	1492	na	1500	680	1500	680	1500	680	abs	680	1500	abs	1500	680	abs	680	no alternative splicing
ERCC6	F689/R1272	971	584	1000	1000	1000	1000	1000	1000	1000	1000	1000	abs	1000	1000	1000	1000	no alternative splicing
cinase (g23)	F85/R913	1285	829	1300	1300	1300	1300	1300	1300	1300	900	1300	900	1300	900	1300	900	alternative splicing specific to bic
	F223/R913	1068	691	1100	1100	1100	1100	1100	1100	abs	700	1100	abs	1100	700	1100	700	
g21	R1/F1	2620	1421		1400		1400		1400		1400		abs		1400		1400	expression confirmed
g39	F291/R873	2747	583		600		600		600		600		600		600		600	no alternative splicing

* lenght of the PCR products were estimated on agarose gel and therefore are not precise as PCR products sequencing

abs for absence of positive PCR, PCR results needed to be confirmed upon repetition

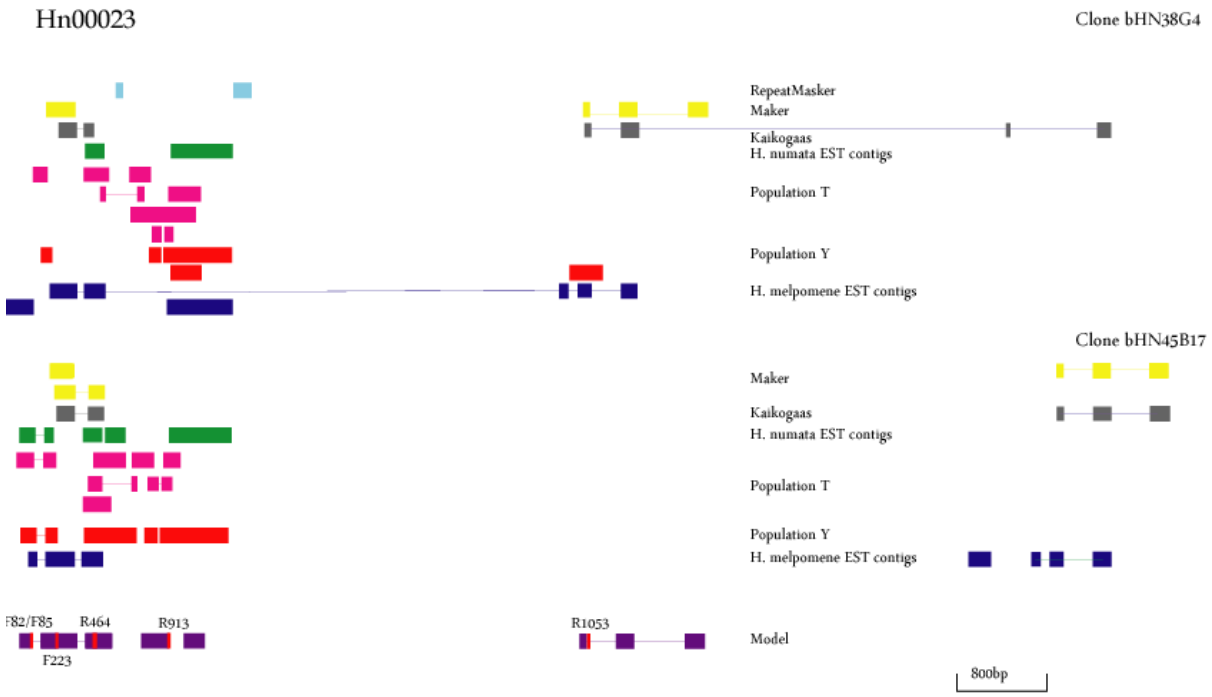
empty boxes stand for PCR on gDNA which were not tested because of the length of the fragment

Models Tested



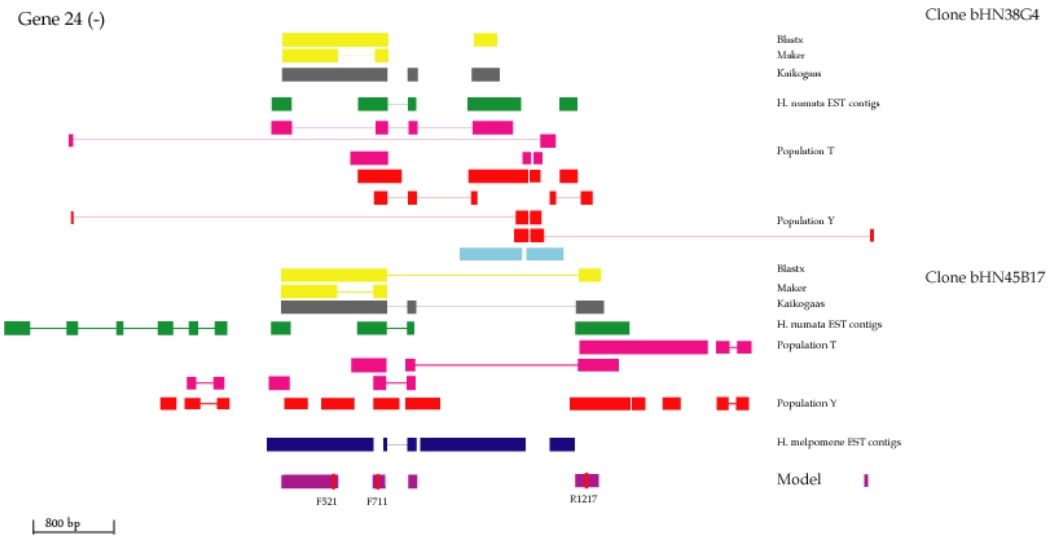
5 - Gene Model for gene 21

Surprisingly, for gene 21 there was amplification in a region that wasn't supported by *H. numata*'s ESTs. This confirms what was predicted by the automated pipeline. The length of cDNA amplified using primers F1/R1 seem to agree with what was expected based on the alignments of the primers in the BAC sequence (aprox. 1400 bp). There was no difference in the amplified fragment length between different races.



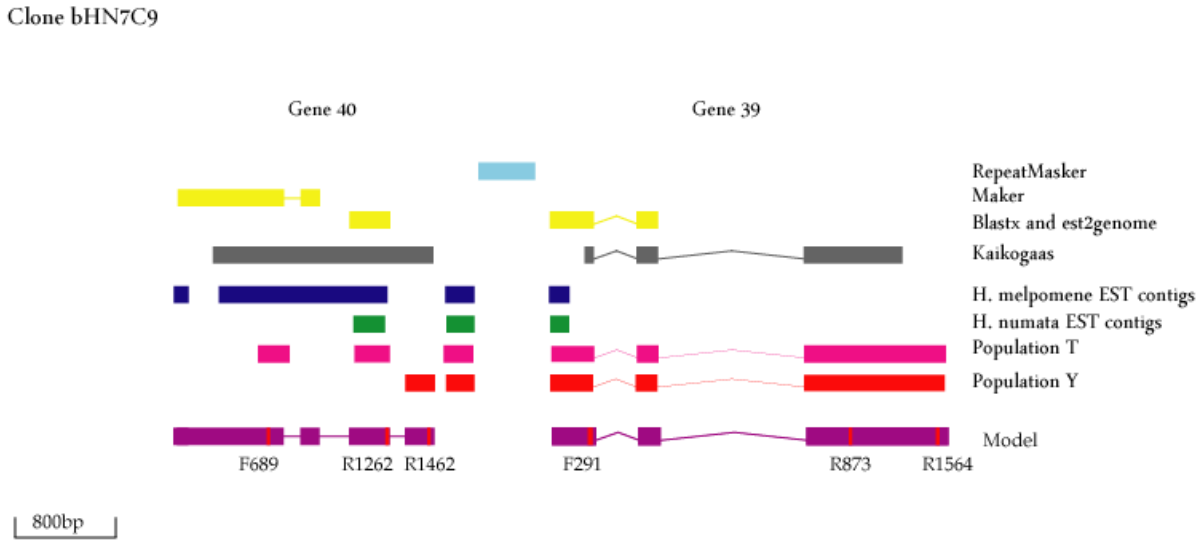
6 - Gene Model for gene 23

It was not possible to test for the presence of the second feature of gene 23 predicted by Maker (primer R1053) because the fragment is too long to amplify with the PCR conditions used in this project. Using the pair of primers F85/R913, the fragment length of the fragments amplified in *H.n silvana* and *H.n. tarapotensis* corresponds to what was expected (aprox. 1300 *vs* 900) validating the model. That is also true for the other pair of primers (F223/R913). However, for the dominant for *H. n. bicoloratus* the cDNA had the same length as the gDNA which may mean the retaining of introns #2 and #3 in my model. This suggests differential splicing between the races but should be confirmed upon sequencing.



7 - Gene Model for gene 24

The length of the fragments amplified seems to confirm my model in both pairs of primers used, validating the features included in the model, namely exons #2 and #3 and introns #1 to #3



8 **- Gene Model for genes 39 and 40**

In gene 39 gDNA amplified fragment length presents a difference of 700 bp relative to my model but the cDNA fragment length corresponds roughly to the same as expected (aprox. 600bp). This means that the intronic region of the gene is larger than predicted.

In gene 40 there seems to be no splicing (amplified gDNA and cDNA have the same length). This challenges the reality of introns #1 and #2 in my model but it is not surprising considering the Kaikogaas prediction and the *H. melpomene* ESTs.

Before taking any conclusion regarding splicing of the genes or the coverage of the EST reads, sequencing of the fragments is necessary to better evaluate the real length of the fragments amplified and this should be accomplished upon sequencing the PCR results

Conclusion and Discussion:

The genes investigated here are of interest because the are located in the breakpoints of inversion in the genomic sequence where Locus P is. As mentioned, it is currently believed that rearrangements in this region could be locking together this set of genes. This locus is indeed inherited in a block but there is still no data about which particular genes responsible for determining wing patterning.

Heliconius melpomene gene	Heliconius numata BAC	Putative name	Domains	Best *B. mori* blastp hit	Best *D. melanogaster* blastp hit
HM00021	Hm11J7	HM00021	ResIII	BGIBMGA005657	NS
HM00022	Hm11J7	Enoyl-CoA hydratase	ECH	BGIBMGA005656	CG6543
HM00023	bHN38G4/bHN45B17	ATP binding protein	DUF265 domain	BGIBMGA005557/BGIBMGA005558	CG10581
HM00024	bHN38G4/bHN45B17	LRR	LRR domains	BGIBMGA005655	Sur-8
HM00025	bHN38G4/bHN14K13	HM00025	WD repeats	BGIBMGA005652	cort
HM00039	bHN7C9	Licorne	Pkinase_Tyr	BGIBMGA005641	Licorne
HM00040	bHN7C9	DNA excision repair protein ERCC-6	Helicase C	BGIBMGA005640	Hel89B
HM00041	bHN45B17	Penguin	PUM_repeats, pumulio RNA-binding domains	BGIBMGA005637/ BGIBMGA005638	Penguin
HM00053	bHN6M17	Lethal (2) Giant Larvae	WD repeats	BGIBMGA005570	Lethal (2) Gant Larvae
HM00054	bHN6M17	Zinc finger protein	Zn finger repeats	BGIBMGA005571	Crooked legs

Of particular immediate interest are:

Gene 41

This gene is similar to *D. melanogaster's* penguin. Experimental evidence supports its involvement in apposition of dorsal and ventral imaginal disc-derived wing surfaces. Also, electronic pipeline suggests RNA binding molecular supporting a role in gene regulation. This is a good candidate since it is involved in wing surface morphology. Search in BioGRID (a database of protein and Genetic Interactions) revealed interaction with P115 with 46 other proteins including ash2 which is associated to imaginal disc-derived wing vein specification and DAAM, dishevelled associated activator of morphogenesis.

Gene 53

This gene is similar to *D. melanogaster's* lethal (2) giant larvae, a protein coding gene which molecular function is myosin II binding. There is experimental evidence for 17 unique biological process terms, including: anatomical structure development; macromolecule localization; reproductive process in a multicellular organism; gland morphogenesis; organelle organization; organ morphogenesis; cell fate determination; basal protein localization; regulation of neurogenesis; actin filament-based process; establishment or maintenance of polarity of follicular epithelium; cellular localization; negative regulation of Notch signalling pathway. There are several described phenotypes in a variety of structures including imaginal discs in different developmental stages.

Gene 54

This gene is similar to the gene crooked legs from *Drosophila melanogaster*. Sequence similarity supports RNA polymerase II transcription factor activity. There is experimental evidence that it is involved in negative regulation of transcription; cell adhesion; negative regulation of Wnt receptor signalling pathway; regulation of transcription from RNA polymerase II promoter; imaginal disc-derived wing morphogenesis; positive regulation of mitotic cell cycle. Dorsal thoracic disc, appendage segment and wing pouch are part of the structures used to annotate the reported phenotypes.

These genes seem to be strong candidates since they are involved in regulation of Notch and Wnt pathways.

- The **Notch** pathway is a cell signalling system highly conserved in most multicellular organisms, which involves gene regulation mechanisms that control multiple cell

differentiation processes during embryonic and adult life. Ligand proteins binding to the extracellular domain inducing cleavage and release of an intracellular domain, which enters the cell nucleus to alter gene expression.The receptor is triggered by direct cell-to-cell contact in a way that groups of cells influence one another to make large structures(citar).

- The **Wnt** pathway involves a large number of proteins that can regulate the production of Wnt signalling molecules and their interactions with receptors on target cells. This network of proteins is well known for its role in embryogenesis. Originally, *wingless* was identified as a recessive mutation affecting wing and haltere development in Drosophila melanogaster but later on it was characterized as a segment polarity gene that functions during embryogenesis and adult limb formation during metamorphosis. Wnts produced from specific sites, such as the edge of the developing fly wing are distributed throughout adjacent tissues in a gradient fashion (morphogenetic signalling). This pathway becomes activated to different degrees in cells of these tissues depending on gene product concentration (and difusion), leading to subtle but crucial differences in genes regulation(citar).

As for the remaining genes, despite the fact their putative roles don't seem to be directly linked to wing development, they should not be discarded from the list of possible candidates because they seem to code for protein domains that can also take part in gene regulation:

Gene 23

With a good hit for ATP binding protein, this gene was reported by Ferguson *et al* 10 as presenting different splice variants, some of them coding. It should be interesting to address if these correlate with different patterns in different populations. The corresponding gene in *Bombyx mori* is BGIBMGA005557. The best hit for *D. melanogaster* presents a protein coding gene with nucleoside-triphosphatase and tranferase activity; e*rgo* its putative ATP binding function can be involved in a variety of biological processes.

Gene 24

The best hit in *D. melanogaster* is Sur-8, a protein coding gene predicted to have Ras GTPase binding function. This is consistent with a possible role in signal transduction, and is also adjacent to gene 23, a probable noncoding transcript with complex patterns of alternative splicing that may have a regulatory function (15)

Gene 25

On previous work in *Heliconius melpomene,* splice variants were identified between closely related races on either side of a phenotypic hybrid zone (24), differing only in the presence of the yellow hindwing bar. Also, two races producing the yellow hindwing bar express at least at least three major isoforms of *HM00025,* whereas two races without the bar produce no splice variants in these regions. This suggests that indeed this gene can play a role in specifying phenotypes and also, that alternative splicing is important in wing pattern specification (15)

Best hit from *D. melanogaster* is a protein coding gene for which there is experimental evidence for anaphase-promoting complex binding. It is involved in female meiosis I; activation of anaphase-promoting complex activity during meiotic cell cycle; female meiosis II; pole cell formation; cyclin catabolic process; female meiosis; cellularization; egg activation.

Gene39

The best hit for Drosophila is protein coding gene for which there is experimental evidence of MAP kinase activity, meaning it is possibly involved in oocyte axis specification; transforming growth factor beta receptor signalling pathway; activation of MAPK activity; mucosal immune response. The phenotypes of the reported alleles were annotated with embryonic segment, egg and egg chamber, pole cell, and dorsal appendage, which suggests an indirect role embryological development.

Gene 40

The best hit in *D. melanogaster* is protein coding gene Helicase 89B, which function is, based on sequence similarity, ATP-dependent DNA helicase activity. There is experimental evidence that it is involved in innate immune response, Toll signalling pathway. It doesn't seem to be directly associated with wing patterning but it may have an important generalized role DNA repair.

For genes **21** (hypothetical protein with a type III restriction domain) and **22** (Enoyl-Coa Hydratase Complexed With Octanoyl-Coa thought to be involved in fatty acid beta-oxidation) data don't seem to support any putative role in patterning in light of what is currently known from other insect species. Nevertheless, convincing evidence for the role of a particular gene in controlling the wing-patterning phenotype requires transgenic work to knock out the genetic factor controlling the wing phenotype, followed by genetic rescue.

These research techniques are not yet routine in Lepidoptera, but we rely on three lines of evidence

to support identification of the genes or factors controlling wing patterning: (1) population genetic data, (2) expression data, and (3) antibody staining .

Here I have described gene models based in an approach that combines manual and automated annotation of the genomic region of interest. I have identified possible splice variation which may be related to pattern variation. However time constrains didn't allow practical analysis of all the genes. In any case this work sets a basis for further work and confirmation by sequencing that with the advance in techniques and genomic tool to explore this system will allow from a more developmental point of view.

Joron, Papa, Beltrán, Chamberlain, Mavárez, *et al. (*2006) hypothesise that this switch in the patterning process is most likely a transcription factor with a number of cis regulatory elements that respond to the spatial information. This transcription factors would trigger a response acting upstream pathways to affect pigment deposition and scale morphology characteristic of each wing pattern element.

There are currently examples of both protein coding and *cis*-regulatory changes playing roles in adaptation and it has been suggested that *certain types* of phenotypic change are generated by *certain types* of mechanisms Stern and Orgogozo 2008 report that in comparison to physiological traits, there are many more instances of morphological evolution being driven by *cis-regulatory* changes than by protein coding mutations. Previous studies show introduction of new exons or inclusion of introns occurring at a higher rate than nucleotide substitution or gene duplication providing, as pointed out by J. H. Marden in 2006, a more relaxed framework in which to address the question of the rise of phenotypic novelty(23).

Combinations of exon-skipping, retention of intronic regions, alternative 5'and/or 3' ends, alternative initiation sites and alternative poly-adenylation, may provide novelty from pre-existing variation without changing the genomic structure avoiding potential pleiotropic and deleterious effects in coding sequences. Also, gene's modular arrangement of 5' CRE (*cis* regulatory elements) is known to regulate temporal and spatial patterns of expression and are very robust to perturbation which "relaxes" negative selection creating nearly neutral pathways through which novel function could evolve rapidly (27) . Work from Wada *et al.* 2004, Marden *et al*. 2001 show the potential of AS in generating evolutionarily phenotypic diversity at intra, interspecific and macroevolutionary scales.

Considering the practical results of this project, indeed the PCR results for gene 23 (kinase) using two pairs of primers (F85/R913 and F223/R913) seem to suggest differential splicing between the dominant form *H.n.bicoloratus*, and *H.n.tarapotensis* and the recessive form *H.n.silvana*. In any case, more tests should be pursued in the same genes and in other genes in order to have conclusive results about the frequency and significance of alternative splicing events in the Locus P.

Perspectives

So far, research has focussed on a trio of species encompassing most aspects of colour pattern evolution including geographic diversification, local polymorphism (H. numata), diverging mimetic associations between closely related species (H. melpomene vs. H. numata) as well as convergent phenotypes between distantly related species (H. erato vs. H. melpomene). Advances in genomic resources, including high-resolution maps, BAC libraries, EST scans, and gene chips, are now offering exciting possibilities for comprehensive analyses of colour pattern change in Heliconius.

Since targeted reverse genetics methods aimed at disrupting or enhancing specific gene expression, such as germline transformation and especially RNA interferenc, have been successfully applied to lepidopteran species, being increasingly transferable to diverse species, these represent a promising way to test the involvement of genes in wing pattern phenotypes for Heliconius.

As new candidate loci emerge the challenge will be to perform experimental studies providing a more detailed picture of the networks connecting switch genes to pigment synthesis pathways, and how these change during adaptive radiation.

Integrating our knowledge of different kinds of pattern specification will allow a fuller understanding of pattern evolution and how developmental processes are shaped by selective pressures.

Future work will give a notion of how 'hot' these genomic hotspots of adaptation really are.

The advantages of phylogenetically broad genome coverage are clear, and comparative analysis of diverse genomes will certainly yield important insights into genome evolution and the relationships among branches of the tree of life.

More than accumulating sequence data for comparative analysis, genomic research allows to pursue a complete understanding of how genetic information is translated to produce an organism, and how modifications in genomic composition and organization give rise to diversity.

For future studies should be interesting to determine if AS has been conserved and recruited across lineages of *Heliconius.* Pattern formation involves probably discrete changes in conserved protein coding or regulatory regions (20).

References

1. Akam M. (1995) Hox genes and the evolution of diverse body plans. *Phil. Trans. R. Soc. Lond.* B **349**: 313–319
2. Arthur W. (2002) Intraspecific variation in developmental characters: the origin of evolutionary novelties. *Am. Zool*. 40
3. Beldade P. and Brakefield P. (2002) The Genetics and Evo-Devo of Butterfly Wing Patterns *Nature Reviews*, **vol 3**
4. Beldade P., Brakefield P., and Long, A. (2002) Contribution of Distal-less to quantitative variation in butterfly eyespots. *Nature* **415**: 315– 318
5. Brakefield P., Gates J., Keys D., Kesbeke F, Wijngaarden P., Monteiro A., French V., Carroll S. (1996) Development, plasticity and evolution of butterfly wing patterns. *Nature* **384**:236-242.
6. Brandi L. Cantarel, Ian Korf, Sofia M.C. Robb, *et al*. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes *Genome Res*. **18**: 188-196
7. Briscoe A. and Chittka L. (2001) The Evolution of Color Vision in Insects. *Annual Review of Entomology* **46**: 571-510;
8. Brown K., Benson W. (1974) Adaptive polymorphism associated with multiple Mu¨ llerian mimicry in Heliconius numata. *Biotropica* **6**: 205–228;
9. Carroll S., Gompel N., Prud'homme B., Wittkopp T., Kassner V. (2005) Chance caught on the wing: cis-regulatory evolution and the origins of novelty. *Dev Biol* **283**: 584–584),
10. Carroll S. (2001) Chance and necessity: the evolution of morphological complexity and diversity. *Nature* **409(6823)**:1102-9.
11. Clarke C. and Sheppard P. (1960) The genetics of Papilio dardanus. *Genetics* **45**:, 439–457
12. Cook S., Vernon J., Bateson M. and Guilford T. (1994) Mate choice in the polymorphic African swallowtail butterfly, Papilio dardanus— male-like females may avoid sexual harassment. *Anim. Behav*. **47**: 389–397
13. Coyne J. and Orr H. (1998) The evolutionary genetics of speciation. *Philos. Trans. R. Soc. Lond.* B **353**: 287–305.
14. Cresko W., Amores A., Wilson C., Murphy J., Currey M., *et al*. (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc Natl Acad Sci U S A* **101**: 6050–6055;
15. Ferguson L. *et al*. (2010) Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus *Molecular Ecology* , **19 (Suppl. 1)**: 240–254

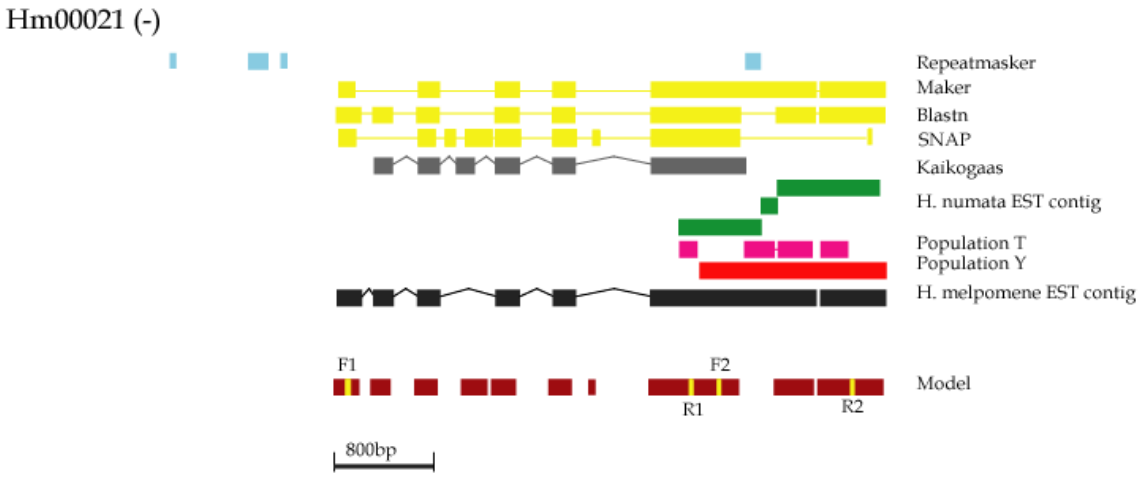
16. Fisher R. (1958) The genetical theory of natural selection versus occasional phenotypic leaps facilitated by conserved developmental pathways, *2nd revised edition. New York: Dover*. (p. 291)
17. Fisher R. 1930. The Genetical Theory of Natural Selection. *Oxford University Press, Oxford.*
18. Galant R., Skeath J., Paddock S., Lewis D., Carroll S. (1998) Expression pattern of a butterfly achaete-scute homolog reveals the homology of butterfly wing scales and insect sensory bristles. *Curr Biol* **8**:807-813
19. Jiggins C., Naisbit R., Coe R. and Mallet J. (2001) Reproductive isolation caused by colour pattern mimicry. *Nature* **411**: 302–305
20. Joron M., Jiggins C., Papanicolaou A. and McMillan W. (2006) Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity *Heredity* **97**: 157–167
21. Joron M., Wynne I., Lamas G., Mallet J. (1999) Variable selection and the coexistence of multiple mimetic forms of the butterfly Heliconius numata. *Evol Ecol* **13**: 721–754
22. Kapan D. (2001) Three-butterfly system provides a field test of mullerian mimicry. *Nature* **409**: 338–340
23. Lynch M. and Conery J. (2003) The origins of genome complexity. *Science*. **302(5649)**:1401-4.
24. Mallet J. (1989) The genetics of warning colour in Peruvian hybrid zones of Heliconius erato and H. melpomene. *Proc R Soc Lond B Biol Sci* **236**: 163–185
25. Mallet J. and Barton N. (1989) Strong natural selection in a warning-color hybrid zone. *Evolution* **43**: 421–431
26. Mallet J. and Joron M. (1999) Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Syst*. **30**: 201–233
27. Modrek B, Lee C. (2002) A genomic view of alternative splicing. *Nat Genet.* **30(1):**13-9
28. Monteiro A, Brakefield PM, French V: Butterfly eyespots: the genetics and development of the color rings. Evolution Int J Org Evolution 1997, 51:1207-1216.;
29. Nijhout H (1991) The development and evolution of butterfly wing patterns. *Washington D. C. Smithsonian Institution Press.* (297 p.)
30. Orr, H., and Coyne, J. (1992) The genetics of adaptation: a reassessment. *Am. Nat.* **140**: 725–742.
31. Payre F, Stern DL (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* **448**:U586-U587
32. Reed R., Serfas M. (2004) Butterfly wing pattern evolution is associated with changes in a Notch / Distal-less temporal pattern formation process. *Curr Biol* **14**:1159-1166

33. Richardson M. and Brakefield P. (2003) Hotspots for evolution *Nature| News and Reviews* **vol 424**
34. Schluter D., Clifford E., Nemethy M., McKinnon J. (2004) Parallel evolution and inheritance of quantitative traits. *Am Nat* **163**: 809–822
35. Stern, D. 1998. A role of Ultrabithorax in morphological differences between Drosophila species. *Nature* **396**: 463–466.
36. Stern,D. (2000) Perspective: Evolutionary Developmental Biology And The Problem Of Variation *Evolution* **54**: 1079–1091

Appendix

Gene Models

Because of the fact that there are no available BAC libraries for the region where *H. numata*'s genes 21 and 22 might be, I used the *H. melpomene*'s clone **11J7** where they were previously described by Ferguson *et al.* 2010 and align *H. numata'*s ESTs against it. This shouldn't imply bias given the high sinteny between both species.

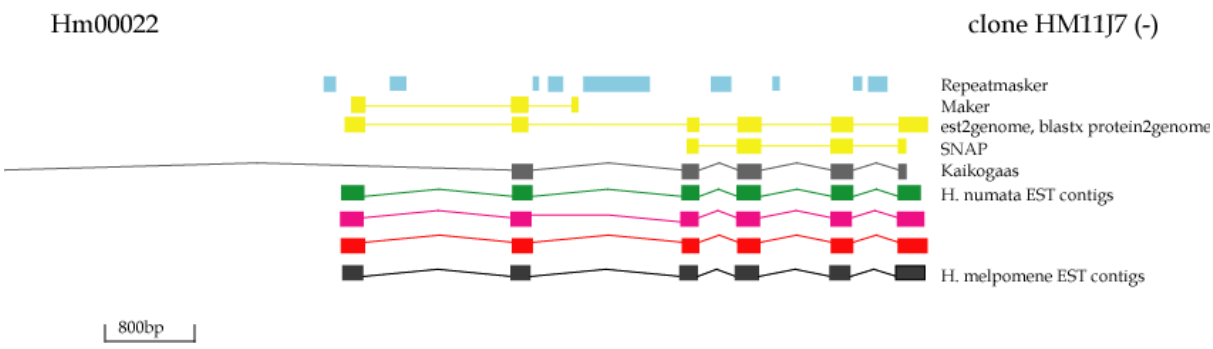


Gene 21

Kaikogaas annotation for this *H. melpomene*'s gene suggests estimated size of 1611bp and includes six putative exons which single hit is a type III restriction enzyme's domain.

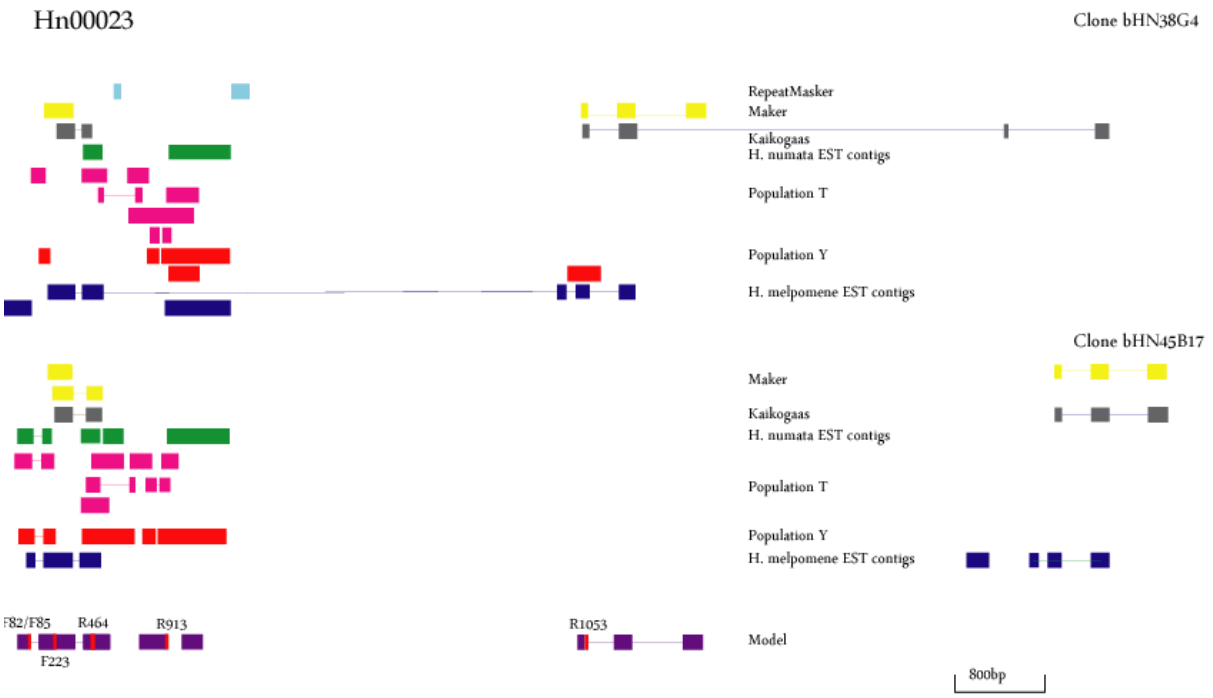
The Maker pipeline gave 3 different predictions.

H. numata's ESTs supported mainly the last putative exon. *H. melpomene*'s ESTs support the automated predictions. When blasted, these features have a good hit for a putative *D. melanogaster* protein whose predicted activity is glutaminyl-peptide cyclotransferase activity.



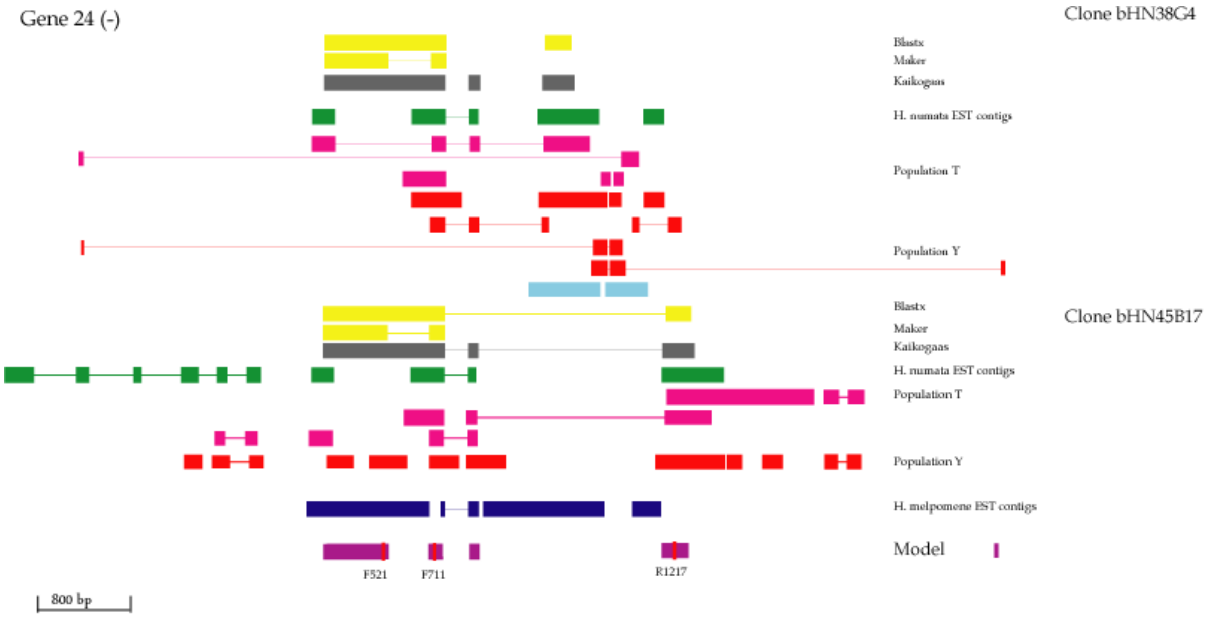
Gene 22:

The Kaikogaas predicts a gene 792bp long with 6 putative exons. The predicted function of this gene is putative Enoyl-Coa Hydratase Complexed with Octanoyl-Coa. There are 3 different Maker predictions, overlapping partially all of them supported by Kaikogaas' feature and by EST sequences. Contigs and raw reads present an exon-intron like organization supporting exactly the features predicted by Maker's first prediction and ending in a polyA tail. Since the support given by the ESTs is clear and confirms the automated models there is nothing to test so this gene was not amplified.



Gene 23:

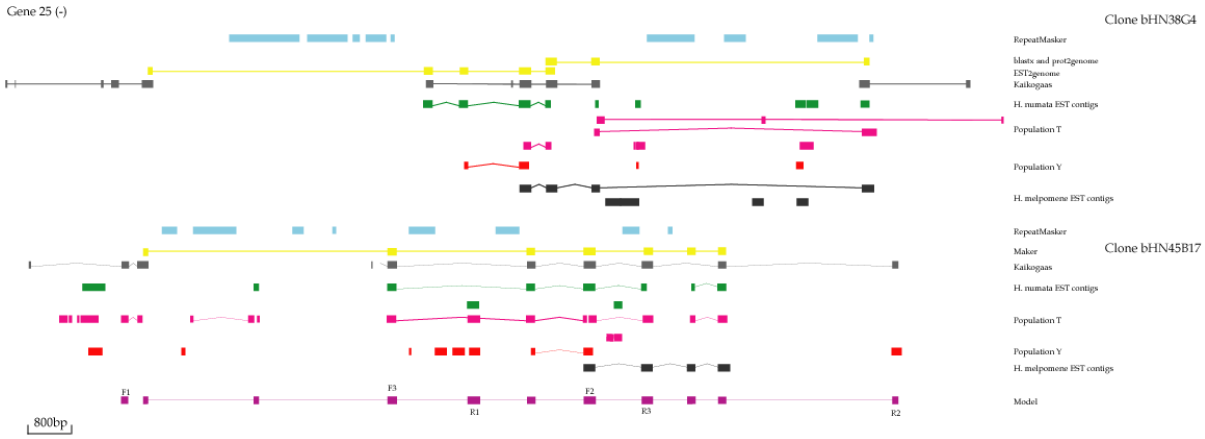
In both clones, the Maker prediction is supported by Kaikogaas prediction. This feature presents one DUF265 domain and consists in a good hit for ATP-binding protein. In both clones there is a second Maker feature also a good hit for ATP-binding protein and supported by a *H. melpomene* contigs. This can be due to the fact that both features are part of the same gene but because of the long length between them the predictors fail to interpret it as such, or, since there is no support from *H. numata* ESTs (only from *H. melpomene*), this feature could be specific to *H. melpomene* (maybe as a result of duplication for instance).



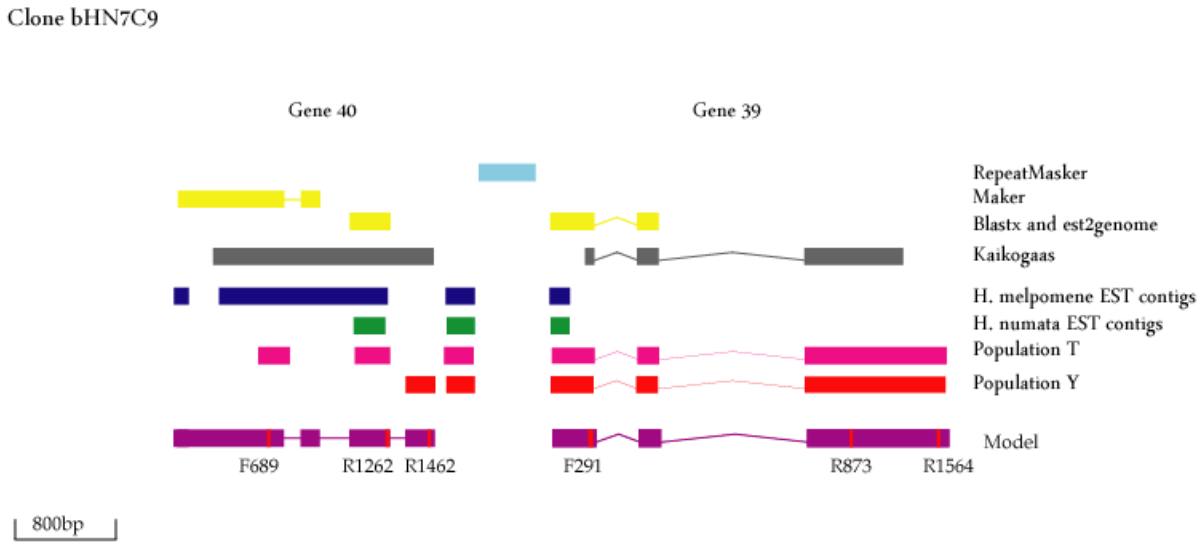
Gene 24

Presenting a good hit for LRR domain, two different Maker predictions overlap suggesting two exons supported by kaikogaas predictor. The vast majority of the raw reads has a PolyA sequence after the third putative exon but some of them continue in an exon-intron like organization suggesting a AS event. For the clone bHN45B17 there's an additional feature supported by *H. numata* ESTs that expands in a region not considered as part of the putative gene. In any case it is worth to test such a feature since it is specific of one clone.

Gene 25



In clone bHN38G4 the Maker pipeline suggests a gene 435 bp long with three putative exons. The est2genome prediction differs completely from this, suggesting four putative exons (765 bp) the last of which aligning with the first one of Maker. The central exons of both features are supported by *H. numata* ESTs but reads align mainly on regions masked by repeat masker. The Maker predictions are confusing and the ESTs support mainly intronic regions (specially those masked). There are discrepancies between population T and Y ESTs, something worth testing In Clone 14K13 the predictors are easier to interpret and Maker supports seven putative exons supported by blastx, protein2genome and est2genome all of which are supported by Kaikogaas. Most of the raw reads are masker by repeatmasker in repetition regions (possibly transposons).

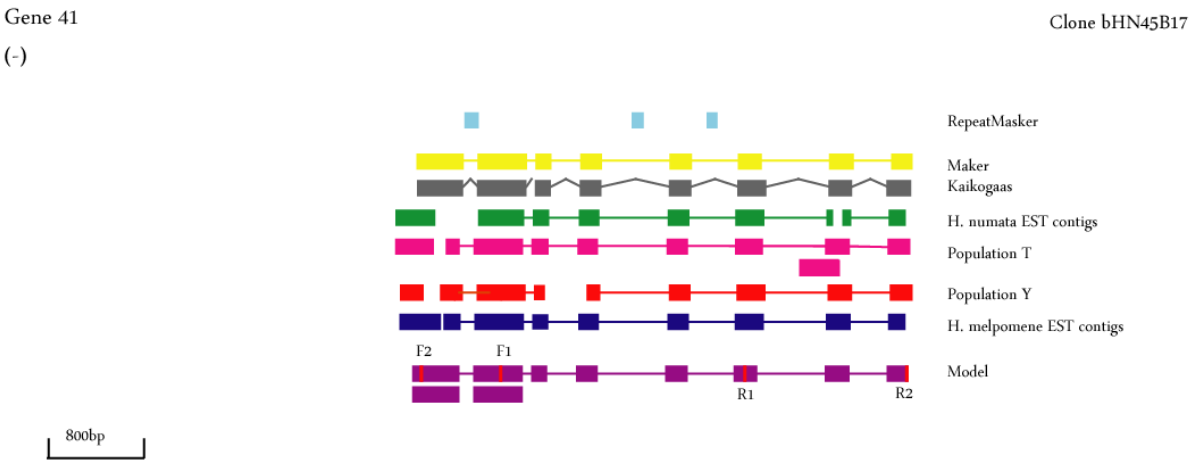


Gene39

In Clone7C9, Maker suggests two exons (506 bp) supported by the EST data available. The Kaikogaas predictor supports these, suggesting one more exon. These features are also supported by *H. melpomene*'s data presenting a good hit for mitogen activated protein kinase. The Raw Reads from population Y support 3 putative exons and finish in a polyA tail.

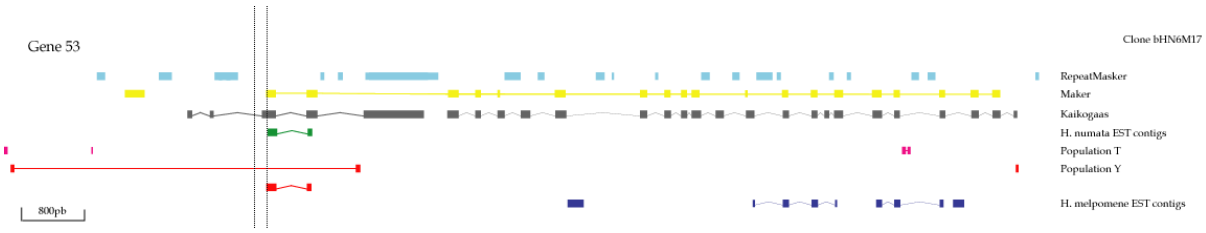
Gene 40

In the same clone and very close to gene 39, the Maker pipeline predicts two putative exons supported by blastx, blastn and est2genome. Kaikogaas overlaps this prediction suggesting one more exon further supported by *H. numata*'s raw reads. The kaikogaas prediction has a good hit for *Bombyx mori*'s gene BGIIBMGA005640 but the support from the ESTs is not spread through the whole length of this feature, which suggest that maybe the gene has two smaller exons instead of a larger one. This gene has a good hit for DNA excision repair protein ERCC-6;



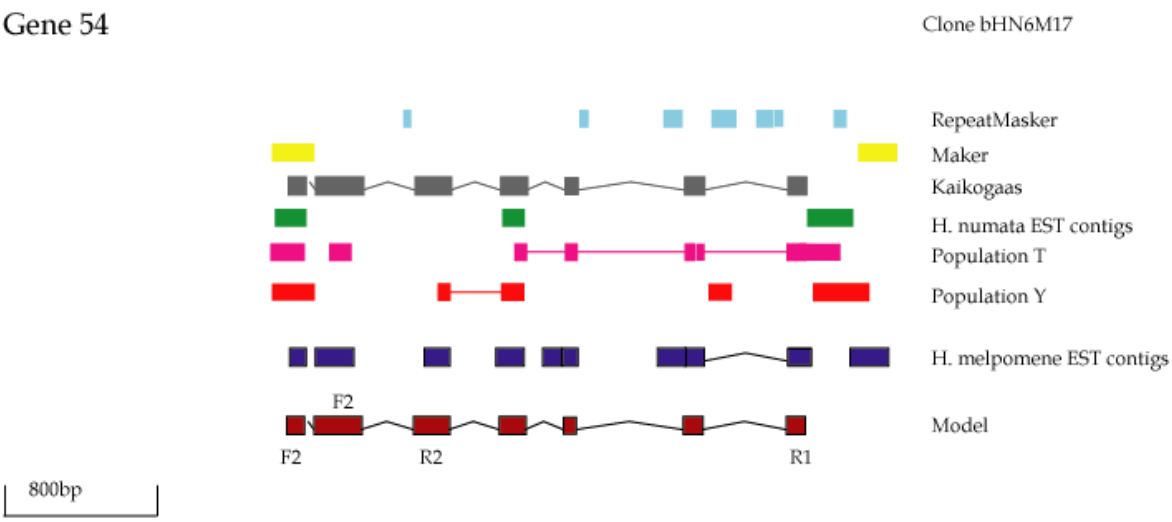
This gene has a good hit for penguin protein and the support from both population of *H.numata* as well as from *H. melpomene* agree clearly with each other suggesting a gene with 8 exons (1878bp). The fact that all the raw reads aligning in the exon #8 present poly adenilation tails suggests that the model is robust regarding the length of the gene. The only features that suggest diffently are some population Y raw reads that present a poly A tail after exon #2 which could be part of other gene (24/LRR) located very close in the clone. Also, the frontier between exon #2 and #3 should be tested for this population because, contrarily to what we can see in population T, there are no raw reads supporting it (i.e. aligning against the putative exons with a gap in the putative intronic region).

Gene 53



Gene 53 has a good hit for Lethal (2) Giant Larvae. Maker predicts 19 putative exons (2319 bp), all of which are supported by one Kaikogaas prediction. A small feature predicted by Maker that didn't seem to be part of the gene has a good hit for the same protein but is not supported by any EST, aligning in a region that could be considered as intronic. This can mean that the gene extends even further than what the predictions suggests. The support from the ESTs for this gene is extremely poor maybe due to the fact that the exons are so numerous and small and, therefore, difficult to find in a blast search or because of the expression levels being low (and there for the gene is not well represented in the 454 EST data).

Most raw reads align in a region masked that has a good hit for transposase protein, meaning it is probably a transposable element. Besides the fact that few EST sequences came as an outcome from the blast search against the genomic sequence, most alignments are not convincing and align in repetition regions. Only *H. melpomene's* sequences seem to support more convincingly some of the putative exons. Since there is so little support from EST data, there is not enough information to produce testable hypothesis and therefore *in vivo* tests were not planned for this gene.



This gene has a good hit for zinc finger protein. The only Maker prediction that has a good hit for zinc finger protein, supported by est2genome and blastn, is 321 bp long and constitutes a single putative exon that seems to be interrupted by stop codons. When looking at the Maker prediction for other insects there's a different scenario: a prediction of 7 putative exons further supported by Kaikogaas prediction (1320 bp). The support from *H. numata*'s ESTs is scarce differing between both populations._Y raw reads support the first putative exon, and the frontier between third and fourth putative exons. T raw reads support both fist and second putative exons not overlapping its whole length, and the putative exons #4, #5 and #6. Most of the sequences were discarded for not having a good alignment or for aligning in intronic regions.

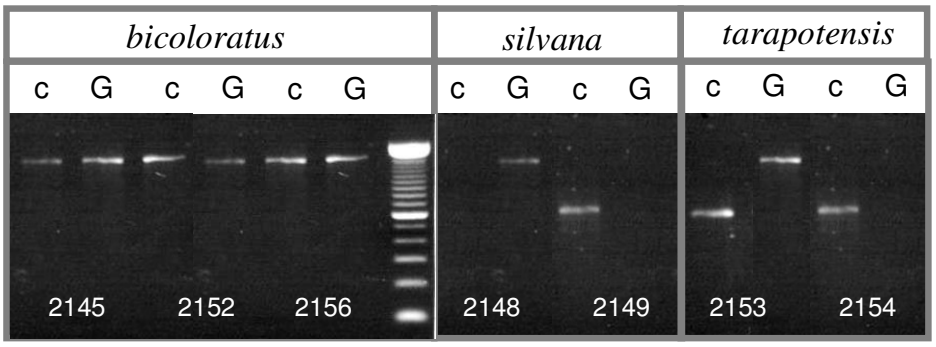
Primers' List

Primer	Sequence	Tm(ºC)
Hn21_HP_F1	AAACCAGCGTATCGGTGTTC	56
Hn21_HP_R1	ATATCTTGTGCGGCATTTCC	54
Hn21_HP_F2	TGCAGATTCAAAACCGACTG	54
Hn21_HP_R2	GCGATGGAGCTGTAGACGAT	57
Hn23_Kinase_F124	GAGACCCTGGTGTCGGTAAA	57
Hn23_Kinase_F223	ATAAAATGCGGGAAGGGTTT	51
Hn23_Kinase_F82	TCAGCAATAATGATCCAAAACATC	51
Hn23_Kinase_F85	GCAATAATGATCCAAAACATCAAA	50
Hn23_Kinase_R1053	CCTGTTTTGCATGGAGTCCT	55
Hn23_Kinase_R464	CCTTTTCTAGAAGGTATTGTGGCTA	55
Hn23_Kinase_R913	TGTTCTGTCTGTTTAGGGGCTA	55
Hn24_LRR_F521	AGTCCATTTGTTTGGGGACA	53
Hn24_LRR_F711	TGGCTTATCAAGCCTTCCTG	55
Hn24_LRR_R1217	TGAAGGGCCTCTATGAGGAA	55
Hn24_LRR_R1514	GCCCACCCAGGACATAGATT	57
Hn25_fizz_F1	CGCAACGTTATCGCCTAGAT	56
Hn25_fizz_R1	GGGTCACTCACTCATCACGA	57
Hn25_fizz_F2	GCTGGCGTCATTTAGTGAGA	56
Hn25_fizz_R2	TTCATTTCTTCGCTCTGGTG	54
Hn25_fizz_F3	TCAATGCTGCAAGTTTGACC	54
Hn25_fizz_R3	GCCAGTCCTATGGCTGAAGA	57
Hn39_Licorne_F291	ACCTCCTCCTCCAGGCTTTA	57
Hn39_Licorne_R1564	ATGGACGATGCGGTCATTAT	53
Hn39_Licorne_R873	ACCAGAAATGCCGAAATCAC	53
Hn40_ERCC6_F689	CTGGCAAGATGTCGGTAGTG	57
Hn40_ERCC6_R1272	CTGTTGGCGTTTCAAGGTTT	53
Hn40_ERCC6_R1462	TGCAATTTTAATCGCTTCCA	50
Hn41_pen_F1	TGGTGCAAATGATGCAATCT	52
Hn41_pen_R1	ATGAAACTGGGATGGAATGC	54
Hn41_pen_F2	GCGTGGAATCTGAGACATCA	56
Hn41_pen_R2	TCCCTGTGATGACTGTTGCT	56
Hn54_zfp_F1	GGAGGAAAATGATTCGCAAG	54
Hn54_zfp_R1	CGTGAGCTGTCATCTCGTGT	57
Hn54_zfp_F2	CAGTGGCAGTGGAGAAGGAT	57
Hn54_zdp_R2	CGCACTTGTGTGGTTTGTCT	56

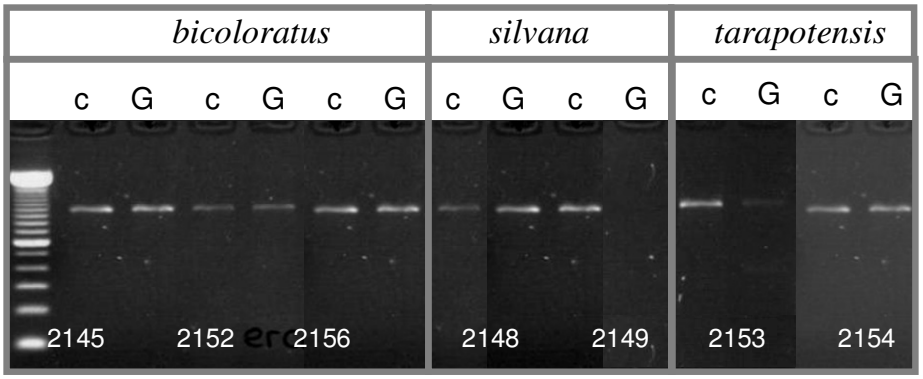
Gene	F- Primer	R-Primer	Model genomic	cDNA	MEGA Alng 11J7mel		To Test
	Hn21_HP_F1	Hn21_HP_R1	2550	1419	2762		ex4 and ex7
21	Hn21_HP_F1	Hn21_HP_R2	3915	2683	4057		in8 and in9
	Hn21_HP_F2	Hn21_HP_R2	1091	1060	1090		in8
					38G4	**45B17**	
	F82	R464	589	383		1580	ex2 and ex3
	F82	R913	1294	832		1288	ex4 and ex5 (end of the gene)
	F82	R1053	4965	972	704	9175	splice variant - 2nd prediction
	F85	R464	595	380	neg	589	ex2 and ex3
	F85	R913	1291	829	neg	1285	ex4 and ex5 (end of the gene)
23	F85	R1053	4962	696	701	9172	splice variant - 2nd prediction
	F223	R464	369	242	367	364	ex2 and ex3
	F223	R913	1065	691	1061	1060	ex4
	F223	R1053	4736	831	4730	8947	splice variant - 2nd prediction
					38G4	**45B17**	
	F521	R1217	1466	697	1482	2553	putative in1, ex2
	F521	R1514	1763	994	1778	2849	Ex 2 and ex4
24	F711	R1217	1276	507	1292	2363	Ex2
	F711	R1514	1573	804	1588	2659	ex4
					bNH38G4	**bHN14K13**	
	Hn25_fizz_F1	Hn25_fizz_R1	6544	762	24315	6544	ex1, ex3, ex4
	Hn25_fizz_F1	Hn25_fizz_R2	14282	1709	15947	14277	ex3, ex5, ex11
	Hn25_fizz_F1	Hn25_fizz_R3	9703	1246	8133	9702	ex3, ex5, in7
25	Hn25_fizz_F2	Hn25_fizz_R2	5758	746	8419	5753	in8, ex11
	Hn25_fizz_F2	Hn25_fizz_R3	1179	283	605	1178	in7
	Hn25_fizz_F3	Hn25_fizz_R1	1552	246	18470	1552	in4
	Hn25_fizz_F3	Hn25_fizz_R2	9290	1193	10102	9285	ex5, in7, ex11
	Hn25_fizz_F3	Hn25_fizz_R3	4711	730	2288	4710	ex5, in5, in7
					7C9		
	F291	R1564	2747	1274	2785		in1, ex 2, in2
39	F291	R873	2077	583	2094		ex3, ex5, ex11
	F689	R1272	969	584	968		put.in1 and ex2
40	F689	R1462	1297	774	1297		Ex 4
					bHN45B17		
	Hn41_pen_F1	Hn41_pen_R1	2069	757	2068		polyA ex2
	Hn41_pen_F1	Hn41_pen_R2	3430	1223	3429		in2,in7
41	Hn41_pen_F2	Hn41_pen_R1	2731	1291	2730		ex1, in2
	Hn41_pen_F2	Hn41_pen_R2	4092	1756	4091		ex1, polyAex2, in7, ex8
					bHN6M17		
	Hn54_zfp_F1	Hn54_zfp_R1	3487	917	3486		in2, ex4, ex5
	Hn54_zfp_F1	Hn54_zfp_R2	1319	515	1318		in2, ex3
54	Hn54_zfp_F2	Hn54_zfp_R1	3911	1341	3923		in1, in2
	Hn54_zfp_F2	Hn54_zfp_R2	1743	939	1755		in1, ex2, in2

PCR results

LRR F521/R1217



ercc6 F689/R1272



Kinase F223/R913