4

Discussion

4.1 Motivation for this study

Two sets of results obtained within our research group prompted this study. The first was an observation that when using representational difference analysis (RDA) to compare highly similar genomes for DNA differences between them, chloroplast fragments, especially the ribosomal subunits, located in the nuclear genome were isolated as RDA products from various plant species. These results indicated that the chloroplast insertions in the nuclear genome were highly variable within different plants of the same species or cultivar and even between different individual plants obtained from the same tissue culture process. It was therefore decided to do a detailed analysis of the plastid insertions in the nuclear genome to better understand the dynamics of nuclear located plastid insertions. The second observation was the isolation of a 215 bp sequence from tobacco with high similarity to a *Bacillus cereus* sequence. Subsequent analysis identified similar fragments within the genome of the grass *Monocymbium ceressiiforme* as well as white clover, *Trifolium repens.* Submitted manuscripts incorporating these results have been met with criticism that the fragments were most likely artifacts of DNA contamination rather than genuine differences between the genomes being compared. It was therefore decided to asses the rice nuclear genome for the presence of all sequences similar to any part of the completely sequenced microbes with which rice was likely to have been in intimate contact, rather than just focusing on genes.

4.2 Discussion of techniques and results

4.2.1 Search techniques and parameters

The first objective of this study was to establish and refine the techniques and parameters used with which these comparisons were done by comparing these results to existing reports of chloroplast insertions in the rice genome.

In this study the raw alignment data for the chloroplast shows a combined length of homology of over 1 Kb in the nuclear geome, this however include a number of repeats. Eliminating all the repeats as well as fragments shorter than 100 bp and with homology of less than 95% or E-values larger than 10^{-20} the combined length was 0.78 Kb. This is less than the 0.9 Kb obtained by Matsuo *et al*., (2005) but since they used E-values of $<10^{-10}$ with unspecified homology, with

the higher stringency used in this study this is to be expected. The reason for the higher stringency was to distinguish between fragments from possible mitochondrial or chloroplast origin, as well as to eliminate older smaller and older fragments. This was achieved when looking at the small amount of co-alignment shown in figure 3.11.

Overall the search results indicate that the methods and parameters used to do the homology analysis was effective and could be used for the further analysis of viral, bacterial and fungal insertions.

4.2.2 DNA homology between the plastid genomes and the rice nuclear genome

This study confirmed the presence of large amounts of plastid DNA in the nuclear genome of rice. While other studies have reported on the nuclear location of plastid DNA in various eukaryotic species, it was mostly reports of specific fragments or more general findings. Matsuo *et al*., 2005 was the first to publish a genome wide analysis of chloroplast insertions. This study provides the first comparative analysis of both the chloroplast and mitochondrial genome insertions in the nuclear genome of any plant species. It is then interesting to find that there is a marked difference in the representation of these two plastids in the nuclear genome of rice. This leads to a hypothesis that firstly the process of transfer and insertion or secondly that the process of elimination of the inserted fragments differs between the chloroplast and mitochondria.

To examine the hypothesis that there is a possible difference in the transfer and insertion of the chloroplast and mitochondria to the nuclear genome and rice, one has to consider studies on transfer rates. Recently, two research groups experimentally detected plastid–nuclear DNA transfer in tobacco, with an apparent transfer rate of a marker gene of 1 per 16,000 pollen grains (Huang et al., 2003) or 1 per 5 million leaf cells (Stegemann et al., 2003). Interestingly, the number of cells in a mature tobacco leaf (Possingham, 1980) is at least 10 times higher than the average number of leaf cells required to select one chloroplast gene transfer event, which indicates that cells within a single leaf are not genetically identical but may differ in their nuclear genome with respect to the pattern of chloroplast DNA integrations. Taking this transfer rate into account, a given gene locus on the tobacco plastid genome is expected to translocate to the nucleus 63 x 10^{-6} times per plant generation or 0.2x10^{-6} times per somatic cell division. Matsuo

et al., (2005) estimate that a given gene locus of the rice chloroplast genome should have transferred to the nucleus at least 4 $x10^{-6}$ times per plant generation. Rice fields generally produce 100 to 300 $x10^{6}$ grains/hectare. Therefore, in a rice field of 1 hectare, there may be hundreds of grains in which a given gene of the chloroplast genome has newly transferred to the nucleus. The occurrence of grains in which the nuclear genome had newly integrated any part of the chloroplast genome would be far higher. The rate of transfer for mitochondrial DNA was determined in yeast by Thorsness and Fox (1990) and was similar to the rates of chloroplast DNA transfer in tobacco leaf cells found by Stegemann *et al*., (2003) at $\approx 2 \times 10^{-5}$ per cell per generation.

While the total length of insertions for both the chloroplast and mitochondria were similar in this study (1070 fragments totaling 779 Kb for the chloroplast and 1329 fragments totaling 614 Kb) their genomes show a remarkable difference in the way they are represented in the nuclear genome (see figures 3.2 and 3.6), with the chloroplast genome represented on average approximately 5.8 times and that of the mitochondria is only represented on average less than 2.1 times. While most of the chloroplast genome is present to some extent, only parts of the mitochondrial genome are present in the nuclear genome, most of the areas of the mitochondrial genome present in high frequencies are the NADH subunits that also account for most of the shared homology with the chloroplast. To differentiate between NADH copies from the chloroplast and the mitochondrion, fragments were assigned to one or the other by identification of the flanking sequences as either chloroplast of mitochondrial. Two copies of the ATPase β-subunit were found in the nuclear genome. The first is located between two mitochondrial insertions while the second is located well outside of any plastid DNA insertions, (see figure 3.9). One of the distinguishing features of these two copies is eight intron-like sequences that interrupt the coding sequence compared to the chloroplast copies. While there are TATA promoter-like motives in the 5' sequence of these copies no data is available on whether they are indeed expressed. In literature the function ATPase β-subunit has not been shown to have been transferred to the nucleus though these results would seem to indicate that it might be in the process of being transferred.

It has been shown that while plastid DNA is frequently transferred to the nucleus, it is also rapidly eliminated (Matsuo *et al*., 2005). If the insertion rates are indeed similar for chloroplasts and mitochondria in rice as well, then there must be some system operating in the nucleus that preferentially eliminate mitochondrial DNA from the nucleus. Unfortunately there is no evidence

available to support this hypothesis. Comparing the two different representation profiles, that of the chloroplast have a high base representation with a few areas that are represented at a much higher frequency while that of the mitochondria show only a few areas that are highly represented. Assuming that the base in the chloroplast representation is the result of younger insertions that has not yet been through the process of elimination and the areas that are highly represented are the remains of older insertions, similar to the mitochondrial profile. If the chloroplast and Mitochondrial DNA is eliminated at the same rate, then a higher insertion rate of the chloroplast might explain this difference in representation observed.

Taking into account that the rates were estimated in tobacco leaf cells for tobacco and yeast cells for the mitochondria, we should consider that in plants only fragments transferred in the gametophytic cells would stand a chance of being transferred to the next generation, after which it could be incorporated into the general population over time. Therefore what happens in the somatic cells does not necessarily reflect the situation in the gametophytic cells. And unless these somatic cells are involved in vegetatively propagated structures, these insertions will be lost. There is no research available to support a hypothesis for a different transfer rate in gemetophytic cells, we might speculate about this from other data. It has been proposed by Richly and Leister, (2004) that because extranuclear DNA is usually maternally inherited in flowering plants. We also know from observations in tobacco by Yu and Russell (1993) that while the chloroplast content decreased to an average of 0.48 plastids/generative cell mitochondria were present in larger numbers at approximately 80 mitochondria per generative cell. Thus if we assume a similar situation in rice, coupled with the hypothesis of Richly and Leister that organelle-to-nucleus transfer of DNA should preferentially occur during the degradation of plastids during pollen formation, the higher decrease of chloroplast should result in a larger amount of free chloroplast DNA in the cell available for transfer to the nucleus. Although this cannot be proven in this study, this would explain the higher abundance of chloroplast DNA in the nucleus observed in this study.

Another contributing factor that might explain this difference is that if insertion happens mainly as whole genome insertions as postulated by Matsuo *et al*., (2005). We know from studies on transformation of *Escherichia coli* with plasmids by Hanahan (1983) that the transformation efficiency declines linearly with increasing plasmid size. Therefore the smaller genome of the rice chloroplast (130 Kb) would be more efficiently transferred and incorporated into the nucleus than the much larger genome of the rice mitochondrion (490 Kb). Although Matsuo *et al*., (2005)

view the process of insertion to occur mainly through the insertion of whole genome fragments or fairly large fragments with the subsequent deletion and fragmentation of these pieces, it is also reasonable to consider that it might occur through multiple insertions of small fragments, some of which might act like transposable elements. A fragment on chromosome 12 of the rice genome shows a position where possible multiple independent insertions of chloroplast fragments took place (see figure 3.4). These insertions are interspersed with short highly repetitive T-rich DNA sequences. The non-contiguous nature of the fragments seems to indicate multiple independent insertions rather than insertion with subsequent deletion of parts of the fragment.

It also appears that the insertions are not randomly distributed over the whole nuclear genome, but rather that they tend to be concentrated in specific areas. There are at least three possible explanations for this: (1) Insertions occur throughout the genome but some regions are "swept' more effectively than others. (2) Some regions of the genome are more available/ receptive than others to insertion events and therefore these regions have a higher number of insertions. (3) Insertions in other areas are mostly deleterious or decrease the cells fitness to such an extent that they are selected against.

Furthermore the size distribution of the insertions shows that the majority of fragments are between 100 – 599 bps long. If we consider that most insertions occur as whole genomes of as large pieces, this indicates that the insertions we are observing for both the chloroplast and mitochondrion is fairly old and has been degraded significantly, but that they are efficiency with which they are degraded decrease significantly once they reach this size class. Combined with this, the insertion of fragments might also occur most efficiently for this size class.

4.2.3 Sequence similarities between the viral genomes and the rice nuclear genome

In this analysis we confirmed a previous report that the Rice tungro bacilliform virus (RTBV) is integrated into the rice nuclear genome. Nagano *et al.,* (2000) reported a highly variable region on chromosome 6 with homology (54 – 59%) to RTBV that was also highly variable in copy number between rice cultivars. Kunii *et al*., (2004) reported on 29 segments homologous to RTBV. Previous studies focused on the characterization of specific fragments through PCR, sequencing and Southern Blot Analysis rather than a whole genome comparison. Using this

approach this study identified additional fragments that have not yet been reported. Up to 40 copies of some regions of the Rice tungro bacilliform virus were found to be present in the rice genome, indicating either multiple integrations or subsequent duplication of some regions. If indeed the RTBV genome has been inserted more than 40 times over the course of time, it would seem that it has been accompanied by extensive elimination. Even so most of the genome is still represented in the rice nuclear genome at various copy levels. It is interesting that we find similar fragments of RTBV only in one other plant species namely *Vitis vinifera*. We could find no reports of RTBV infecting *Vitis*, but is seems likely that it does or that there is a as yet unidentified/ un-sequenced counterpart of RTBV associated with *Vitis*.

Furthermore this study provides evidence of possible integration events between rice and five different RNA viruses and the rice genome were also identified. Fifty-one fragments that show similarity to rice related RNA viruses have been identified. The integration of RNA viruses into the plant genome has not yet been reported (Harper *et al*., 2002), except for the finding that grapevine genomic DNA carries the entire gene of a potyviral coat protein (CP) and the potyviral 3′UTR. Potyviral-homologous sequences were also found in tobacco DNA, albeit in a rearranged form (Tanne and Sela, 2005). The sequence fragments identified in this study show between 40 – 70% identity with the viral genome and range in length between 184 and 1800 bps. This variation in homology would seem to indicate different times of integration during the evolution of the rice genome. Only fragments of the viral genomes are present. Since many of the RNA viruses of rice have genomes containing several fragments rather than circular genomes, partial integration would be more likely. With animal and bacterial retroviruses the early steps of replication involve reverse transcription of the viral RNA genome to make a cDNA copy followed by the integration of that cDNA copy into a chromosome of the host cell. The integration reaction requires specific sequences at the ends of the viral cDNA, which bind the viral-encoded integrase and other proteins to form pre-integration complexes (Schröder *et al*., 2002). The infection cycle of plant viruses is not known to include an integration event, mainly because integration of retrovial DNA is facilitaded by a virally encoded integrase (Patience *et al*., 1997). Plant pararetroviruses generally lack the gene for this enzyme, and integration is not required for virus replication (Jakowitsch *et al*., 1999), even so there are is evidence of the integration of non-retroviral genomes into plant genomes (Tanne and Sela, 2005; Staginnus and Richert-Pöggeler, 2006).

With the absence of an integrase enzyme the integration of viral sequences into the plant genome must involve a recombination event. In the absence of viral sequence in the host genome, recombination must be non-homologous (Jakowitsch *et al*., 1999). If there are already viral sequences already present, the recombination could be homologous. It is still unknown if the complete viral DNA genome itself integrates or whether integration occurs through replication intermediates such as cDNAs. During replication, pararetrovirus genomes form multiple copies of nuclear mini-chromosomes and gapped dsDNA replication intermediates, and the ssDNA geminivirus genomes are replicated to high levels in the nucleus. Both modes of replication provide potential recombinogenic sequences for integration, via illegitimate recombination, in cells undergoing active genetic processes (Hull *et al*., 2000). Integrated sequences of banana streak virus and RTBV integrated sequences suggest that the cDNAs can integrate as well (Harper et al., 1999; Ndowora et al., 1999; Jakowitsch et al., 1999; Kunii et al., 2004). It is however possible that a retroelement integrase function can be used *in trans*, as has been suggested for SINE element integration from LINE element integrase (Eikbush, 1992). It might also be plausible that RNA virus segments are integrated after the viral segments get reverse-transcribed by reversetranscriptase used by other viruses in the cell.

It would seem likely that viral sequence integration can occur during every infection. However, as most viruses do not infect meristematic tissues, the integrations are not usually fixed and are lost on passage through seed. Furthermore cells and plants in which functional copies of the integrated virus can readily excise, be transcribed or otherwise result in disease might not be viable. Rapid rearrangement and/or deletion of viral sequences through recombination would effectively disrupt the process. This is a possible reason for all reported cases of integration involving rearrangement or deletions.

The possibility that viral DNA might insert regularly into plant genomes, has considerable implications for plant genome evolution. Little is yet known about the contribution of integrated viral sequences to plant genome organization, function and evolution. Similar to vertebrate endogenous retroviruses (Patience et al., 1997), integrated pararetroviral DNA could act as insertional mutagens; could contribute strong constitutive promoters to neighboring plant genes, altering gene expression patterns; or they could accumulate to generate new repetitive sequence families. As components of the genome, they can be altered, recombined and amplified, providing another source of variation.

Overall this analysis presents the first comprehensive assessment of viral integration and contribution to the rice nuclear genome.

4.4.4 Sequence similarities between the bacterial genomes and the rice nuclear genome

With the exception of *Agrobacterium tumefaciens*, transfer of DNA from bacteria to plants remains a controversial subject. Recent research by Broothaerts *et al*. (2005) has shown that several bacterial species outside the *Agrobacterium* genus, modified with a Ti-plasmid were able to facilitate the transfer of foreign DNA to plants. We know that various eukaryotic genes are of bacterial origin, presumably acquired during endosymbiosis and subsequent transfer of DNA from the pre-chloroplast and pre-mitocondria (Timmis *et al.*, 2004).

This analysis aimed to identify recent lateral DNA transfer events from bacterial genomes to that of the rice genome. In the first whole genome comparison between a bacterium and rice, done with the *Bacillus subtilus* genome, various sequences with a significant degree of similarity were identified. On closer analysis of these sequences the similarity were found to be mostly between the 16S and 23S rDNA genes of the chloroplast located in the nuclear genome of rice. This is not a surprising observation since the chloroplasts are thought to have originated from cyanobacteria (McFadden, 2001). When cross similarities between the bacteria especially within the rRNA regions were eliminated as possible HGT events then the possible contributions from the various organisms became more differentiated with a ten-fold greater amount of similarity between rice and *Bacillus* than either of the other two bacterial species. These similarities are unlikely to be simply due to chance events or more regions would have been expected to be identified with the other organisms and/or regions of the microbial genomes.

The 770 bp fragment found exclusively between *Bacillus* and rice genome (excluding the rRNA fragments), which also had the highest similarity of the fragments identified between rice and *Bacillus* with sequence similarity with part of the HD domain protein from *Bacillus*. This region is within the known gene in the rice genome. This gene family includes members from a wide variety of eukaryotes. It is interesting to note that the similarity with this fragment starts in an intron, includes 2 exons and then terminates in a longer intron. Using Blastn on the NCBI website the two top results (100%) is in rice and *Bacillus* followed by alignments 99% and less in various organisms, including Arabidopsis thaliana. This sequence seems to represent a good

candidate of a transfer event of DNA from bacteria to plants. The transfer could have occurred before the split between monocots and angiosperms, thus its presence in Arabidopsis as well as maize and rice. Its absence from other plants can be due to lack of sequence information, or can point either to multiple transfer events or the loss of that sequences from these genomes.

The sequence similarities found between the rice genome and *Pseudomonas* included a possible candidate for horizontal acquisition namely the 181 bp fragment (1:29540592-29540767) that showed sequence similarity to a Glycine cleavage system P-protein from *Pseudomonas syringae.* Since the fragment falls within an intergenic region of about 80,000bp and is not part of a functional gene in rice the sequence similarity is unlikely to be as a result of sequence conservation during the course of evolution. It is also possible that this might have been part of a larger sequence fragment that has been degraded to its current form as is the case for inserted fragments of the mitochondrion and chloroplast.

Xanthomonas might have been expected to have the highest number of similarities due to its intimate association with rice. However, since *Xanthomonas* is a rice pathogen, the association mostly results in the death of the infected cells. Further, the interaction occurs mostly on leaf tissue not contributing to the formation of seeds. Therefore, any DNA transferred during the interaction is unlikely to be incorporated in the progeny of the plant. In contrast, *Bacillus* has a beneficial relationship with the plant by suppressing the growth of pathogens. This provides a greater chance of stable integration of transferred DNA into the progeny and this expectation is borne out by the higher degree of homology found between rice and *Bacillus.* The similarities identified here are unlikely to be simply due to chance events or more regions would have been expected to be identified with the other organisms and/or regions of the microbial genomes.

From literature as well as this study we can see that there are a lot of sequences in plants with similarity to bacterial sequences that cannot be explained by conventional evolutionary theories, and where horizontal gene transfer seems to offer the best explanation. Though it is assumed to be rare events, the data for transfer of chloroplast and mitochondrial DNA to the nucleus demonstrate that though it happens quite frequently in leaf tissue the actual incorporation into the general plant population is much lower since only DNA incorporated into the nucleus of cells that will give rise to progeny will stand a chance of being incorporated into the plant population. A similar scenario will be true for horizontal transfer of bacterial DNA. While it might happen much more often in stems and leaves we might never observe these. Another important factor

will be the specific association of the bacterium and the plant, since pathogens usually kill the cells they invade or the cells are killed by a hypersensitive response in the plant to prevent further infection, any transfer of DNA will be lost. Gene transfer between bacteria in a symbiotic association with the plant such as *Rhizobacteria* in roots could be much higher but since roots are seldom the progenitors of new offspring these transfers would not be transferred to subsequent generations.

This analysis demonstrates that a whole genome comparative analysis can be useful in identifying similarities between genomes that can be attributed to horizontal DNA transfer events between unrelated genomes.

4.4.5 Sequence similarities between the *Magnaporthe* genome and the rice nuclear genome

The whole genome comparison between rice and *Magnaporthe* revealed 144 sequences, 100 bp or longer, with significant similarity in the rice genome. These sequence fragments are spread over the rice genome (as shown in figure 3.17). Ninety one percent of the of the alignments were between conserved regions of genic sequences such as beta-tubulin and polyubiquitin while nine percent were either in unknown or *Magnaporthe* related sequences. There were also a number of fragments identified as putative heat-shock proteins while other fragments could not be identified. When the rice sequences identified as similar were used in the reverse blast, these similarities were greatest between rice or monocots and *Magnaporthe*, but there was also wide homology across different fungi and plants. The higher amount of sequence similarity found between *Magnaporthe* and rice can be attributed to the higher number of conserved genes between fungi and plants. However, a number of examples of DNA sequences were found with the only significant homology being between rice and *Magnaporthe grisea*, of either unknown function in both organisms or annotated only in *Magnaporthe* as hypothetical proteins. Some insertions showed significant homology only to *Magnaporthe grisea* such as the 131 bp insertion on chromosome 5 (5775966 - 5776097) to locus XM_370434.1 of *Magnaporthe grisea* ($9e^{-27}$; 87% identical without gaps), this falls within the TIGR Locus LOC_Os05g10630.1 (putative O-sialoglycoprotein endopeptidase). Another such fragment was a 109 bp sequence on chromosome 4 (23876766-23876857) to locus XM_365128.1 ($9e^{-11}$; 83% identical without any gaps) although the latter of these was eliminated from the final statistics of similarity due to the high e-value. Another fragment on rice chromosome 6: 4668751-4669004

shows significant sequence similarity only in maize (AY104186.1; 8e-63) and the fungus *Coprinopsis cinerea okayama* (XM_001833442.1; 4e-31). In rice this forms part of TIGR Locus LOC_Os06g09290.1 (putative 26S protease regulatory subunit 7). This mitigates against evolutionarily-related genes being the only source of these genomic similarities. The non-genic conserved fragments are most likely to represent concrete examples of horizontally transferred DNA between *Magnaporthe* and rice, but each would require in-depth analysis.

Analysis of the fragments to determine the relatedness between the rice and *Magnaporthe* sequences were done by using similar fungal and plant sequences identified using Blastn homology searches in the plant and fungal databases to do multiple sequence alignments and draw sequence dendograms. Since these sequences are mostly less than a 1000 bp long a significant phylogenetic analysis was not possible. Using these as guides some conclusions can be drawn about the origin and relationships between the fragments found to be similar between rice and *Magnaporthe*. The first set of dendograms (figures 3.18 - 20) would seem to indicate conserved sequences inherited *via* a common ancestor since the plant and fungi group separately. *Magnaporthe* is the closest grouping fungi to the plants which could be explained by its close evolution with the cereals as a cereal pathogen.

The second set of dendograms (figures 3.21 – 5.24) supports horizontal DNA transfer events between *Magnaporthe* and rice or a common ancestor of maize and rice (figure 3.23). The first dendogram in this set (figure 3.21) represents a 253 bp sequence only identifiable as a hypothetical protein in rice, but with similar sequences in different plants and fungi present. It is clear however that the sequence from rice groups well away from the other plants with that of the fungi as would be expected for a horizontal transferred sequence. The second dendogram in this set (figure 3.24) represents a 124 bp hypothetical protein sequence from rice and the dendogram again suggest a transfer event rather than homology via a common ancestor. In this dendogram the fungus *Blastocladiella emersonii* group between the plants. It is thought that *Blastocladiella emersonii* which is an aquatic fungus of the Chytridiomycete class, and lying at the base of the fungal phylogenetic tree could have retained some ancestral characteristics of fungi and animals or fungi and that were lost in members of late-diverging fungal species (Ribichich et al., 2006). The third dendogram (figure 3.23) also represents a sequence that could only be classified as a hypothetical protein in rice of 100 bp. The dendogram would again support a transfer event from an early ancestor of *Magnaporthe grisea*, *Aspergillus terreus* and

Neurospora crassa to an early ancestor of both rice and maize. The fourth dendogram (figure 3.24) was constructed using a 645 bp sequence annotated as *rub1 mRNA for polyubiquitin*. This dendogram also shows that the sequence from the i*ndica* cultivar group of rice groups with the fungi and separate from the sequence in the *japonica* cultivar group, indicating a DNA transfer event that took place after the two cultivars split an estimated 0.4 mya (Zhu and Ge, 2005).

The third set of dendograms support a hypothesis for multiple DNA transfer events between plants and fungi. The firs dendogram (figure 3.25) that were drawn using a 109 bp fragment identified as a hypothetical protein from *Magnaporthe grisea*, groups *Magnaporthe* and rice together while sorghum and maize are grouped with the fungus *Chaetomium globosum* an ascomycete found on grasses with known antifungal activity against the *Magnaporthe grisea* and wheat leaf rust *Puccinia recondita* (Kim *et al*., 2005). The other two cereals wheat and barley grouped separately. Cotton (*Gossypium hirsutum)* grouped with the fungus *Ustilago maydis* as well as Arabidopsis and grape (*Vitis vinifera*). The second dendogram in this group (figure 3.26) that was constructed using a 190 bp fragment identified as a putative heat shock protein in rice, shows the cereal sequences group together while wheat (*Triticum aestivum*) group closer to the fungus *Blastocladiella emersonii* than rice maize and sorghum. The rest of the fungi follow interrupted by *Lilium longiflorum*, *Brassica rapa* subsp. *pekinensis* and tobacco (*Nicotiana tabacum*); the only other plants in which a sequence of significant similarity could be found.

This analysis shows that while some of the sequence similarities between the genome of rice and that of *Magnaporthe* may originate as a result of a common ancestor there are some sequences for which this similarity can only be explained by a transfer event from a fungi (*Magnaporthe grisea* or related fungi) to the rice genome or that of an ancestor of rice and the other cereals. The exact model of how transfer and integration is facilitated is still unclear but illegitimate recombination is a likely explanation.

4.3 Insertion dynamics

This study clearly shows that there are sequence similarities in the rice genome with “non-plant” sequences. For rice genome assembly 80% of the sequences were from paired (forward and reverse) reads with an average clone size of ~1700 bp (18.5-fold genome coverage). More than fivefold coverage was from randomly selected clones, with the remainder from resequencing gaps or low-quality regions. The resulting sequences were analyzed for contamination from non-rice DNA sources (~500,000 reads) or rice repetitive DNA (~1,500,000 reads) (Goff et al., 2002). Since the fragments identified in this study are all relatively short (shorter than the BAC reads ass well as the stringency with which the rice genome has been assembled and the fact that these fragments are flanked by rice-related sequences; it is highly unlikely that they are artifacts of contaminating microorganisms in the original DNA preparations.

Through the various analyses of the insertion sites, it is also clear that the insertion sites of these sequences do not appear to be randomly distributed throughout the genome and has at least three possible explanations:

i. More transfer occurs but some regions are “swept’ more effectively than others due to adverse consequences of exogenous DNA insertions.
ii. Some regions of the genome are more available/ receptive than others to insertion events and therefore these regions have a higher number of insertions or,
iii. Insertions in other areas are mostly deleterious or decrease the cells fitness to such an extent that they are directly selected against.

One view is that only genes that convey a certain benefit will be transferred between organisms (Jain *et al*., 1999; Gogarten and Townsend, 2005) but it is unlikely that benefit could be the determining factor in the actual transfer. Any benefit would rather affect the persistence of the transferred DNA. Alternatively, the degradation and/or rearrangement of any inserted fragment might itself possibly confer a benefit. Therefore it would be more accurate to state that certain genes or regions have a better chance of surviving the mechanisms in the nucleus responsible for degrading insertions and repetitive elements and therefore will be maintained and integrated into the functioning of the organism by conferring a positive selection pressure. Another factor that must be considered is that the position of insertion might play an equal or greater role than its actual coding function, through the influence of the inserted fragment on the adjacent genes and sequences and *vice versa*. It is also clear with the analysis of the organellar insertions that

not all regions of the organellar genomes are present in the nuclear genome at the same frequency. This is especially true for the sequences of the 16S and 23S ribosomal subunits of the chloroplast and the NADH subunits of the mitochondrion. Possible reasons for this are that they either are preferentially inserted, or, more likely, because they are not removed at the same rate as the other insertions. This might be because the genome has an evolutionary built-in mechanism to recognize and protect these essential gene sequences.

4.4 Models for DNA transfer

For the insertion of foreign DNA to take place, the DNA firstly has to enter the nucleus and secondly need to be incorporated into the genome. For the mitochondrion and chloroplast the only obstacle to cross would be the nuclear envelope. While for viruses, bacteria and fungi the cell wall and cell membrane would present an initial and more robust barrier. In host-pathogen interactions, the invading microorganism generally does not invent novel metabolic pathways; instead it insinuates into the existing cellular processes and adapts them for its life cycle. Thus, nuclear uptake of foreign DNA shus as the T-complex from Agrobacterium spp and viral genomes probably follows one of the pathways for nuclear transport of cellular RNAs.

Transport of nucleic acids through cell membranes is a biological process basic to all living organisms. Nucleic acid molecules are transported through membrane channels during host-pathogen interactions (e.g. transport of viral genomes into the host cell) as well as during normal cellular processes (e.g. nuclear export/import of mRNA) (Citovsky and Zambryski, 1993). Molecular transport across the nuclear envelope involves many different proteins and nucleic acids. This transport is bidirectional and occurs exclusively through the nuclear pore complex (NPC), integrated into the two membranes of the nuclear envelope (Akey, 1992). In the passive state, the NPC allows diffusion of small molecules (up to 40 kDa) (Akey, 1992), while the transport of larger molecules occurs by an active mechanism mediated by specific nuclear localization signal (NLS) sequences contained in the transported molecule (Garcia-Bustos *et al*., 1991). The best studied case of nuclear import of DNA in plants is the *Agrobacterium* T-DNA system. As discussed in the introduction Agrobacterium transforms plant cells through the Ti plasmid. The Ti plasmid has two important genetic components. One, the T-DNA, is copied and transferred to the plant cell and the second component of the Ti plasmid, the virulence (vir) region, provides most of the trans-acting products for T-DNA transfer. In addition to functioning

in nuclear targeting, VirE2 acts as a single-stranded (ss) DNA binding protein (SSB) (Christie, *et al*., 1988; Citovsky *et al*., 1988). The T-strand with a molecule of VirD2 covalently attached to its 5'-end likely exists as a folded and collapsed structure. Following cooperative binding of VirE2, the ssDNA is unfolded to form a long and thin protein-ssDNA T-complex. The T-complex is composed of three structural elements: one copy of the T-strand, one VirD2 molecule, and more than 600 copies of VirE2 (Citovsky *et al*., 1988). The T-strand is not sequence specific, and any DNA sequence located between the 25-bp T-DNA border repeats can be transported to plants and function as T-DNA (Citovsky *et al*., 1992). Thus, the T-strand probably does not possess specific nucleotide sequences for nuclear uptake; instead it likely is passively transported into the nucleus by its associated proteins, leaving the other two components of the T-complex, the VirD2 and VirE2 proteins, to function in nuclear transport (Citovsky and Zambryski, 1993). There may be plant cellular proteins analogous to VirE2 that serve as molecular chaperones coating and unfolding nucleic acids and targeting them to and through nuclear pores. If true, this could explain observations that VirE2 function is not essential on some hosts (Stachel and Nester, 1986); in this case, one can argue that another plant-cell SSB can provide the VirE2-like function. This idea supports the possibility that nuclear transport of Agrobacterium spp. T-complex may represent a generalized process by which ssDNA or RNA molecules move within the cell; i.e. as unfolded nucleic acid--protein complex.

In plants, however, nucleic acids also can be transported between cells. Plant cells are connected by cytoplasmic channels called plasmodesmata (PD) that allow the transfer of nutrients and signals necessary for growth and development. Although plasmodesmata and nuclear pores are structurally different, both are complex proteinaceous pores involved in active bidirectional traffic of macromolecules (Citovsky and Zambryski, 1993). Although essential for plant development and function, plasmodesmata represent an 'Achilles heel' that can be exploited and manipulated by viruses to allow them to spread throughout plant tissues (Oparka and Roberts, 2001). Most plant viruses enter the initially infected cell following mechanical damage inflicted by a biological carrier (insect, fungus, etc) or by abrasion. After initial infection, plant viruses move into adjacent healthy cells through plasmodesmata, the only direct link between plant cells. Some viruses move through plasmodesmata as intact virions, causing permanent modification to plasmodesmal structure. Viruses such as the Cauliflower mosaic virus (CaMV) and the Tomato spotted wilt virus use protein tubules, encoded by viral proteins to pass through the PD (Hull, 1992; Storms *et al*., 1995). Others viruses, such as the Dahlia mosaic virus and Tobacco etch virus (TEV) pass through the PD as intact virions (Santa *et al*.,

1998). Other viruses such as the Tobacco mosaic virus, cause more subtle and often transient alterations to plasmodesmata, allowing the viral genome to move as a ribonucleoprotein complex, and encode for viral MPs which modify plasmodesmata (Citovsky and Zambryski, 1993). The variety of viral movement mechanisms suggests that a unified 'strategy' for viral movement is unlikely to occur, given the variety of viral proteins and genome organizations involved (Roberts and Oparka, 2003).

Though the exact method of DNA transfer between the organelles and the nucleus as well as between the genomes of species remains largely unclear. Transfer of mitochondrial DNA to the nucleus has been inferred to occur by an RNA intermediate (Adams *el al*., 1999), because the nuclear copy resembles an edited mitochondrial mRNA rather than an unedited gene. The exact model for chloroplast insertions is unknown but is thought to occur mainly as complete insertions or in large tracts (Matsuo *et al*., 2005) rather than through an mRNA intermediate and, as proposed in this thesis, to occur during pollen formation. It is possible that DNA from the plastids as well as invading bacteria and fungi are non-specifically bound to movement proteins and transported into the nucleus or to adjacent cells and then to the nucleus where it is incorporated into the genome through non-homologous recombination or even through homologous recombination with similar plant sequences.

4.5 Benefits of insertions

Continuous transfer of DNA from the plastid to the nucleus must either have a neutral effect or confer some sort of positive selective advantage otherwise natural selection would have selected against a phenomena that is undirected. A possible benefit of translocation of genes from the plastids to the nucleus is that genes are moved from the plastid with no recombination and a high redox-load to an environment with recombination and without the associated redox-load of the plastids. This would be beneficial to the genes of the plastids but the continuous addition of more DNA into the nucleus could eventually lead to 'genome obesity' (Bennetzen and Kellogg, 1997). A report by Van der Vyver et al. (publication submitted) has shown that the mutation frequency in the nuclear located plastid sequences is far higher that in other nuclear sequences and that similar mutations occur in the same regions of these nuclear located plastid regions during independent radiation experiments. This would suggest that these 'non-functional' plastid sequences in the nucleus might play an important role as mutation buffers in

the nuclear genome. The incorporation of DNA of microbial origin in the plant nucleus might lead to the acquisition of new genes increasing the fitness of the plant. Inserted tracts of non-coding DNA are probably dealt with in the same way as that of the plastids which is continually eliminated. Acquisition of microbial and viral sequences might also lead to increased resistance against the specific pathogen.

4.3 Contribution of this study to science

It is clear from this study that whole genome comparative analysis can provide valuable information on genome regulation, variability and interactions. This study details some of the dynamics of the nuclear genome of rice and the potentially fluctuating contributions from plastid and mitochondrial genomes, their internal interaction, as well as potential their interaction with microbial genomes resulting in DNA exchange. The starting hypothesis of this study was that the rice genome contains DNA fragments acquired through horizontal DNA transfer events from the genomes of its plastids and mitochondria as well as from microorganism living in association with rice. To evaluate this hypothesis a bioinformatical approach was used to make whole genome comparisons between the rice genome and each of the plastid and mitochondrial genomes, as well as with various completed viral, bacterial and a fungal genome. While insertions of the plastid genomes into the nuclear genome of plants and in particular rice is well documented (Matsuo *et al*., 2005), this study was the first to do a comprehensive comparison of the mitochondrial and chloroplast insertions in the nuclear rice genome regarding the total amount of each in the nucleus, the representation of the different regions of the plastid and mitochondrial genomes in the nucleus and the specific sites of insertion of each within the nuclear genome. This study presented new evidence that showed that there is a difference in the representations of the mitochondrial and chloroplast genomes in the nuclear genome. It is hypothesized that this is either due to the different sizes of the two genomes or to different mechanisms that regulate the insertion and deletion of DNA from the different organelles. was by The data from this study supported the second alternative showing that not all areas of the organellar genomes are equally represented, but that some sequences, like the rDNA sequences from the chloroplast and the NADH sequences from the mitochondrion, are represented at a much higher frequency, consistent with the notion that there are processes of selective insertion or deletion of these transferred sequences.

A further contribution of this study towards the understanding the dynamics of the nuclear plant genome was to investigate the insertion of viral DNA sequences. Overall this study presents the first comprehensive assessment of viral integration and contribution to the rice nuclear genome. The study confirmed a previous report of inserted fragments of the rice tungro bacilliform virus (RTBV) (Nagano *et al.,* 2000) but also identified additional fragments that have not yet been reported. Another important finding was the identification of possible integration events of five different RNA viruses into the rice nuclear genome. Fifty-one fragments that show similarity to rice related RNA viruses were identified. The integration of RNA viruses into the plant genome has not previously been reported. These viruses could form an important source of foreign DNA contribution to the variation in plant genomes that has previously been ignored.

A third leg of this study investigated the possible contribution of bacterial genomes to that of the rice genome. By comparing the genomes of three diverse bacteria with different levels of association with plants, this study was able to identify a number of sequence similarities between the rice and the bacterial genomes. While some of these sequence similarities may be ascribed to conservation of sequences during evolution others point to horizontal transfer events. The methods followed in this study, using different diverse bacteria in the comparison proved valuable in distinguishing between evolutionarily conserved sequences and sequences that are potential candidates for horizontal transfer events.

The fourth part of this study investigated the possible DNA contributions to the plant genome by fungi. Performing a whole genome comparison between the nuclear genome of rice and that of *Magnaporthe grisea* resulted in the identification of 144 sequences that shared significant sequence similarity between the two genomes. Further analysis of these fragments showed that while some of the similarity could be contributed to conserved evolution, others were most likely due to a horizontal transfer event. The events could be the result of transfers either from fungi to plants (and specifically between rice and *Magnaporthe*) or from plants to fungi.

From this study it appears that exogenous DNA insertions might be targeted to certain regions of the chromosome rather than occurring at random. The identification and characterization of these regions might be a powerful tool to develop highly polymorphic markers in rice or other organisms. These regions might be especially valuable in clonal plants with low genetic diversity that would otherwise hinder the use of genetic markers. Furthermore, understanding which areas are more receptable to insertions could be valuable in the application of genetic

transformation of crop plants in targeting gene insertion into certain regions that would ensure stable expression and inheritance of the transgene. Understanding the dynamics between plant nuclei and the organelles as well as between plants and the microbes in their environment in terms of gene transfer is also important in the debate about 'gene escape' from genetically engineered organisms. In plants, one strategy to limit gene escape is so called 'transgene containment' (Daniell *et al*., 1998), whereby the transgene is located in the chloroplast rather than the nucleus. Because chloroplasts are degraded during pollen formation, it is thought that the pollen will therefore be free of the transgene and thereby will prevent gene escape through pollen to non-transgenic plants. As discussed in Chapter 2 of this thesis, degradation of chloroplast during pollen formation might be the preferred stage for transfer of the chloroplast DNA to the plant nucleus and a strategy for 'gene containment' through chloroplast transformation might limit gene escape but is not a failsafe method to prevent it.

Because of the recorded relative high frequency with which mitochondria exchange genes, researchers have focused on chloroplast sequences "which seem essentially immune to HGT" (Bergthorsson *et al*., 2003) to do phylogenetic studies in plants (including DNA barcoding; Chase *et al*., 2005; Rubinoff *et al*., 2006). This study clearly shows that caution is necessary when interpreting plant phylogenies using chloroplast genes as they may not reflect the underlying phylogeny of the organism.

4.6 Future perspectives

The mechanisms whereby horizontal transfer and insertion of foreign DNA take place are still poorly understood. While further bioinformatical comparative studies like this could provide important information on this subject and help to elucidate some of the important factors, they are limited by the fact that we are looking only at a representative genome of any species that might not reflect the diversity of insertions that might be present in a given population of plants or even the diversity within a plant. Another clear need is to uncover very recent transfers that could provide further insight into the transfer process and answer several outstanding questions: how long are the DNA tracts that are transferred; does DNA move back and forth between donors and recipients, or is transfer unidirectional; are there lineages experiencing higher levels of DNA transfer which have common characteristics that might suggest a transfer mechanism? Addressing these questions will require both broad surveys and dense sampling.

To determine or demonstrate horizontal DNA transfer between microbes and plants *in vivo* is difficult, since the major criticism of such studies would be that the observed results are simply a measure of the contamination of the plant DNA sample with the microbe, rather than gene transfer between the two genomes. One possible experiment to address the question would be to co-culture a microbe containing an antibiotic resistance gene, with the appropriate plant-specific signals to be functional in the plant, and a plant cell-suspension or callus culture. Treating the culture with an appropriate antibiotic treatment to kill the microbe it does not have resistance to and subsequently selecting plant cells with antibiotic resistance acquired through transfer of the antibiotic resistance gene from the microbe to the plant would confirm transfer. Further analysis can then determine the exact position of the insertion of this gene into the plant genome and the extent of the transferred DNA. The selection would also indicate the rate at which a single gene is transferred and if regions outside the functional gene were co-transferred then it would also give a frequency for the transfer of any microbial fragment. However this experiment would not address the question of how frequently such a transfer would be maintained in the absence of strong positive selection, an event that would occur most likely at a far lower frequency than would the transfer of 'non-functional' DNA sequences between the plant and microbe. Massive PCR screening and sequence analysis would be necessary to identify the presence of non-functional sequences, and care would have to be taken to eliminate the possibility of genetic contamination of the samples.

Another shortfall of current research in horizontal transfer of DNA is that so far all reported cases of horizontal gene transfer in plants involve evolutionarily distant donors and recipients (Mower *et al*., 2004; Wikström *et al*., 2001; Bergthorsson *et al*., 2004). This is most likely because studies rely on the phylogenetic signal of the donated DNA being discordant with that of the host to confidently detect horizontal transfer events. If transfer events occur within a plant family or between closely related species, gene sequences may not be divergent enough to provide strong enough evidence to support a hypothesis involving horizontal transfer. This limitation may present a significant barrier to obtaining a comprehensive view of the tempo and pattern of plant-to-plant transfer events. If any of the dominant modes of transfer involve mechanisms, such as illegitimate pollination, that favor closely related donors and recipients it would not be detected or accepted using the current criteria needed to support such observations as likely resulting from a horizontal transfer event.