1

Introduction and literature review

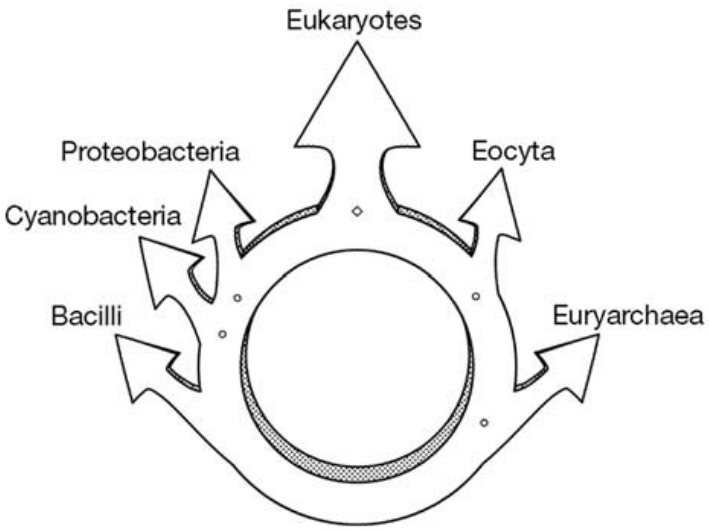
With the onset of whole genome analysis, made possible by the increasing amount of available sequence data, our view on horizontal gene transfer (HGT) is rapidly changing. Scientists are becoming aware that transfer of DNA between non-related species is occurring at much higher rates than was previously thought. This introduction presents an overview of the relevant literature available on this subject, particularly focusing on transfer events involving plant genomes as the recipient genome from various donor genomes including plants, viruses, bacteria and fungi.

1.1 The eukaryotic genome

Genomics and proteomics have greatly increased our awareness of the differences between eukaryotic and prokaryotic cells. The origin of the eukaryotic cell is enigmatic and complex. Early studies of nuclear-encoded enzymes, transfer RNAs, ribosome structures and ribosomal RNA catalogues implied deep, but unresolved, connections between prokaryotes and eukaryotes (Rivera and Lake, 2004). Informational genes (genes involved in transcription and translation) are most closely related to archaeal genes, whereas operational genes (genes involved in cellular metabolic processes, such as amino acid biosynthesis, cell envelope and lipid synthesis), are most closely related to eubacterial genes (Rivera *et al*., 1998). It has been difficult to reconcile these conflicting results with the origin of eukaryotes because of the complicating effects of genome fusions and horizontal gene transfer (HGT) on phylogenetic reconstructions. Several groups have inferred that the eukaryotic nuclear genome derives from HGT through the fusion of archaebacterial and eubacterial genomes (Feng *et al*., 1997; Moreira and Lopez-Garcia, 1998; Rivera *et al*., 1998; Horiike *et al*. 2001; Rivera and Lake, 2004), but this interpretation has recently been called into question (Kurland *et al*., 2006).

Using a recently developed algorithm to do a conditioned reconstruction based on the two character states of gene presence and absence, which can reconstruct genome fusions, Rivera and Lake (2004) presented evidence that the eukaryotic genome arose through the fusion of two diverse prokaryotic genomes. One fusion partner branches from deep within an ancient photosynthetic clade, and the other is related to the archaeal prokaryotes. The eubacterial organism is either a proteobacterium or a member of a larger photosynthetic clade that includes the Cyanobacteria and the Proteobacteria. They therefore suggested that a more accurate representation of the classic tree of life should be a ring of life (Figure 1.1). However, to state

that the eukaryotic cell is the result of a simple fusion of an archaeon and bacterium is an oversimplification. This is supported by the existence of 347 eukaryotic signature proteins (ESPs) identified by Hartman and Federoy (2002). This finding agrees with the predictions of the ABC hypothesis. This hypothesis assumes that the nucleus formed from the endosymbiosis of an archaeon and a bacterium in a third cell, C. One can symbolize this new conjecture as E = A + B + C. This theory would imply that there are three cellular domains despite the large infusion of prokaryotic proteins into the eukaryotic cell because of endosymbiosis (Hartman and Federoy, 2002).



Fugure 1.1: A schematic diagram of the ring of life. The eukaryotes plus the two eukaryotic root organisms (the operational and informational ancestors) comprise the eukaryotic realm. Ancestors defining major groups in the prokaryotic realm are indicated by small circles on the ring. The Archaea, shown on the bottom right, includes the Euryarchaea, the Eocyta and the informational eukaryotic ancestor. The Karyota, shown on the upper right of the ring, includes the Eocyta and the informational eukaryotic ancestor. The upper left circle includes the Proteobacteria and the operational eukaryotic ancestor. The most basal node on the left represents the photosynthetic prokaryotes and the operational eukaryotic ancestor (adapted from Rivera and Lake, 2004).

These hypotheses that attribute eukaryote origins to genome fusion between archaea and bacteria (Feng *et al*., 1997; Ribeiro and Golding, 1998; Lopez-Garcia and Moreira, 1999; Cavalier-Smith 2002; Rivera and Lake, 2004) are uninformative about the emergence of the cellular and genomic signatures of eukaryotes namely the subdivision into subcellular compartments (SSCs) and the eukaryotic signature proteins (ESPs). The presence of ESPs and SSCs are used to suggest an alternative hypothesis: while archea, bacteria and eukaryotes might have shared common ancestors, the eukaryotic subdivision into sub-cellular compartments (SSCs) define a unique cell type that cannot be deconstructed into features inherited directly from archaea and bacteria (Kurland *et al*., 2006). Because their cells appear simpler, prokaryotes have traditionally been considered ancestors of eukaryotes (Knoll, 1992; Baldauf, 2003). Nevertheless, comparative genomics has confirmed a lesson from paleontology: evolution does not proceed monotonically from the simpler to the more complex (Andersson and Kurland, 1998; Klasson and Andersson, 2004; Olsen, 1999). The many ESPs within the subcellular structures of eukaryote cells provide landmarks to track the trajectory of eukaryote genomes from their origins. It is agreed that the unrooted tree of life divides into archaea, bacteria, and eukaryotes (Figure 1.2), whether using gene content, protein-fold families, or RNA sequences (Woese *et al*., 1990; Korbel *et al*., 2002; Snel *et al*., 2002),

On such unrooted trees, the three domains diverge from a population that can be called the last universal common ancestor (LUCA). In this case it is the hypothetical node at which the three domains combine in unrooted trees (Kurland *et al*., 2006). Once the data based on the best-match BLAST protocol are set aside, there seems to be no phylogenetic data available to support the idea that the eukaryotic genome originated as a fusion of bacterial and archaeal genomes. Rather, there are phylogenetic data, such as that for the translation apparatus, the transcription apparatus, and glycolysis, suggesting that all three domains are vertical descendents of a common ancient ancestor (Olsen and Woese, 1996; Canback *et al*., 2002; Woese, 1998; 2000; 2002).

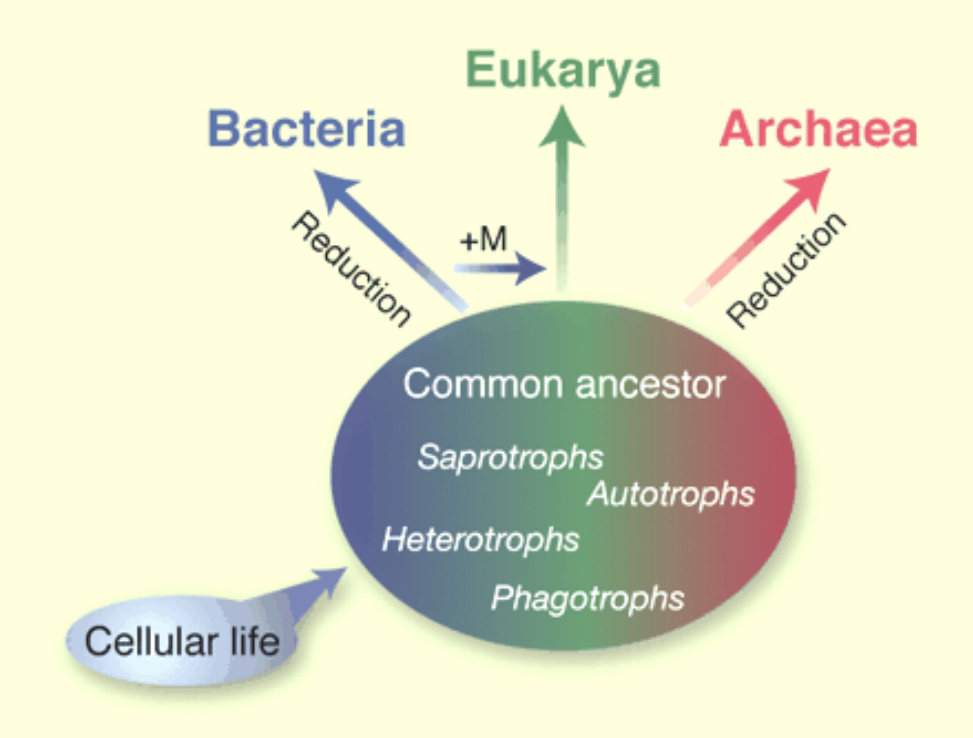


Figure 1.2: The common ancestor of eukaryotes, bacteria, and archaea may have been a community of organisms containing the following: autotrophs that produced organic compounds from CO_2 either photosynthetically or by inorganic chemical reactions; heterotrophs that obtained organics by leakage from other organisms; saprotrophs that absorbed nutrients from decaying organisms; and phagotrophs that were sufficiently complex to envelop and digest prey. +M: endosymbiosis of mitochondrial ancestor. (adapted from Kurland *et al*., 2006).

Regardless of the exact origin of eukaryotic cells, it is believed that they acquired genetic material through horizontal transfer events from various sources and that it is an ongoing process. There is evidence of gene transfer in all lineages of life such as the recently described transfer of genes from bacteria and nematodes to insects (Hotopp *et al*., 2007). The following review however only focuses on the documented transfer events between plants and various microbial donors.

1.2 Origin of the chloroplast and mitochondrion

Chloroplasts and mitochondria are in many aspects similar to each other. Both function to generate metabolic energy and both have their own sets of genes that are more similar to those of prokaryotes than those of eukaryotes and as well as their own protein-synthesizing machinery. It is now commonly believed that both chloroplast and mitochondria are the consequence of endosymbiotic events. The 'Endosymbiotic Theory' was first proposed by former Boston University Biologist Lynn Margulis in the 1960's and officially in her 1981 book "Symbiosis in Cell Evolution". Although now accepted as a well-supported theory, both she and the theory were ridiculed by mainstream biologists for a number of years. Thanks to her persistence, and the large volumes of data that support this hypothesis gathered by her and many other scientists over the last 40 years, biology can now offer a plausible explanation for the evolution of eukaryotes. According to the 'Endosymbiotic Theory', eukaryotes evolved when archaeal and eubacterial cells merged in anaerobic symbiosis. The archaeal cell provided the cytoplasm while the eubacterial cell (a spirochete) allowed for mobility and, eventually, mitosis. Some of these anaerobic cells then incorporated oxygen-respiring eubacteria to become mitochondria-containing aerobes from which most protoctists, animals, and fungi evolved. Finally, some of these aerobes went on to incorporate photosynthesizing cyanobacteria to become chloroplast-containing algae and plants. The divisions or domains implied by this description (Archaea, (true) Bacteria, and Eukarya) are consistent with the widely acknowledged classification system described by Olsen *et al*. (1994).

Mitochondrial genome sizes are variable (Table 1.1) and are unrelated to the complexity of the organism. Most multicellular animals have small mitochondrial genomes with a compact genetic organization, the genes being close together with little space between them. Lower eukaryotes, as well as flowering plants, have larger and less compact mitochondrial genomes, with a number of the genes containing introns. Chloroplast genomes have less variable sizes (Table 1.1). In many eukaryotes the circular genomes coexist in the organelles with linear versions and, in the case of chloroplasts, with smaller circles that contain subcomponents of the genome as a whole. The latter pattern reaches its extreme in the marine algae called dinoflagellates, whose chloroplast genomes are split into many small circles, each containing just a single gene (Zhang *et al*., 1999). We also now realize that the mitochondrial genomes of some microbial eukaryotes (e.g. *Paramecium*, *Chlamydomonas* and several yeasts) are always linear (Nosek *et al.*, 1998).

Table 1.1: Mitochondrial and chloroplast genome sizes

Species	Type of organism	Genome size (kb)
Mitochondrial genomes		
Plasmodium falciparum	Protozoan	6
Mus musculus	Vertebrate (mouse)	16
Homo sapiens	Vertebrate (human)	17
Aspergillus nidulans	Ascomycete fungus	33
Saccharomyces cerevisiae	Yeast	75
Brassica oleracea	Flowering plant	160
Arabidopsis thaliana	Flowering plant	367
Oryza sativa	Flowering plant	490
Zea mays	Flowering plant	570
Cucumis melo	Flowering plant	2500
Chloroplast genomes		
Pinus tunbergii	Conifer	120
Pisum sativum	Flowering plant (pea)	120
Marchantia polymorpha	Liverwort	121
Oryza sativa	Flowering plant (rice)	136
Triticum aestivum	Flowering plant	136
Zea mais	Flowering plant	140
Arabidopsis thaliana	Flowering plant	155
Nicotiana tabacum	Flowering plant (tobacco)	156
Nephroselmis olivacea	Algae	201

Genome sizes obtained from NCBI

1.3 DNA transfer

1.3.1 DNA transfer between the organelles and nucleus

Fragments of chloroplast and mitochondrial DNA are often found in nuclear genomes (Farrelly and Butow 1983; Scott and Timmis 1984; Ayliffe and Timmis 1992; Thorsness and Fox 1990; Sun and Callis, 1993; Yuan *et al*., 2002) and the transfer of DNA from the chloroplast to the nucleus is known to be an ongoing and frequent process (Matsuo *et al*., 2005; Shahmuradov *et al*., 2003; Ayliffe and Timmis, 1992**,** Stegemann *et al*., 2003). Throughout evolution, chloroplasts and mitochondria appear to have lost most of their ancestral genes. It is thought that many genes have been either transferred to the nucleus (i.e. by an 'endosymbiotic gene transfer') or lost completely (Martin and Hermann, 1998; Race *et al*., 1999). During evolution, organelles export their genes to the nucleus, but re-import the products with the help of transit peptides and protein-import machinery, so that proteins are retained in organelles, but most of the genes are not. This process, over time, concentrates genetic material in nuclear chromosomes. Gene-regulatory processes under the control of the nucleus are more complex and interrelated than those under the control of organelles (Herrmann, 1997).

Some chloroplast genes have been extensively studied, especially the rbcL gene for the large subunit of Rubisco (reviewed by Clegg, 1993). In land plants and green algae, the rbcL locus is found on the large single copy region (LSC) in the chloroplast genome, and the rbcS locus for the small subunit of Rubisco is found in the nuclear genome (Clegg *et al*., 1997). However, in cyanobacteria, the rbcL and rbcS genes are adjacent to one another (Nierzwicki-Bauer *et al*., 1984). Thus, the rbcS gene in plants was probably transferred to the nucleus, and subsequently lost, from the chloroplast very early in the evolution of the plants. An example of a more recent transfer is the rpl22 gene in legumes (Gantt *et al*., 1991). Gantt *et al*. (1991) also found that tobacco rpl22 probes hybridized in all angiosperm cpDNA tested, except in legumes. However, rpl22 was confirmed to be in the nucleus, thus legumes has lost rpl22 from the chloroplast genome after its transfer to the nuclear genome. An example from the mitochondrion is the α-subunit of F_1 ATP*ase* which exists in mitochondrial DNA in some eukaryotes but in nuclear DNA in others (Gray, 1992). Furthermore the ribosomal protein gene *rps10* exists in the mitochondrial genome in some angiosperm species, but in the nuclear genome in others (Wischmann and Schuster, 1995; Adams *et al*., 2000). It has also been reported that the respiratory gene *cox2*, which is normally present in mitochondria, is variably involved in the nuclear genome in legume

species. Some legume species possess the gene in both mitochondrial and nuclear genomes, some in the mitochondrial genome only, and others in the nuclear genome only (Adams *et al*., 1999). Gene transfer also takes place between the organelles (Joyce and Gray, 1989; Menaud *et al*., 1998). In Arabidopsis thaliana for example, a gene coding for methionyl-tRNA synthetase in the mitochondrial genome may have originated in the chloroplast (Menaud *et al*. 1998). In wheat, three tRNA mitochondrial genes were found that show high sequence similarity to chloroplast genes (Joyce and Gray, 1989).

1.3.2 DNA transfer between viruses and plants

Unlike animal and bacterial viruses, it was believed that plant viruses integrate rarely, if at all, into their host genomes (Grierson and Covey, 1988). Observations over the past few years have changed this view, as an increasing number of integrated plant viral sequences are being found in plant genomes. Most plant viruses have single-stranded RNA genomes. There are however two groups of DNA viruses that infect plants: the single-stranded DNA *Geminiviridae* that replicate via a rolling circle mechanism and the double-stranded DNA (dsDNA) *Caulimoviridae*, comprising the caulimoviruses and the badnaviruses, that replicate by reverse transcription (Grierson and Covey, 1988). Examples of viral sequences that have been found integrated into plant genomes are a single insertion of sequences related to a geminivirus, which has a single-stranded circular DNA genome, into tobacco nuclear DNA were of multiple direct repeats of partial geminivirus sequence (Kenton *et al*., 1995; Bejarano *et al*., 1996; Ashby *et al*., 1997). These sequences included only the origin of replication and the adjacent viral replication protein, transcription was not detectable and there was no associated virus infection. Examples of integrated viral sequences in plant genomes include:

I. The banana streak virus (BSV)

BSV is a member of the *Badnavirus* genus and the causal agent of viral leaf streak of banana and plantain (Lockhart, 1986). BSV contain a circular dsDNA genome of 7.4 kbp in size (Harper and Hull, 1998). During tissue culture, infections can arise in healthy plants from integrated BSV sequences (Harper *et al*., 1999). Every *Musa* spp. examined to date contains BSV DNA integrated into its DNA. Probing of *Musa* genomic libraries with BSV probes has identified two classes of BSV integrants. The first class of

sequences consists of partial BSV sequences (Geering *et al*., 2001; Ndowora *et al*., 1999). The second class consists of multiple copies of a complete BSV genome, which are believed to be the source of BSV infections that arise during tissue culture (Harper *et al*., 1999; Ndowora *et al*., 1999).

II. Tobacco vein clearing virus (TVCV)

TVCV is a distinct member of the family *Caulimoviridae*, differing from typical caulimoviruses in both genome organization and biological properties. It is a plant pararetrovirus that occurs only in *Nicotiana edwardsonii* (Lockhart *et al*., 2000). The first evidence of TVCV sequences integrated in the *N. edwardsonii* genome was established by hybridization analysis. Cloned TVCV genomic DNA hybridized to Eco R1- and Hind III-restriction digested fragments of *N. edwardsonii* total genomic DNA that were larger than virion genomic DNA, suggesting that the plant DNA moiety hybridizing to TVCV DNA was an integral part of the host genome and not free virion DNA (Lockhart *et al*., 2000). At present, no genomic clone containing the entire TVCV genome has been identified.

The genome of *N. tabacum* also contains pararetrovirus-like sequences, tobacco pararetrovirus-like (TPV-L) (Jakowitsch *et al*., 1999). Despite a copy number of thousands in the *N. tabacum* genome, there was no evidence of any functionally intact viral sequences, the TPV-L clones sequenced containing frameshifts and stop codons. No episomal infections were detected, although there was evidence that the sequences were transcribed. Based on an analysis of the junctions between plant and viral sequences, it was proposed that integration occurred by illegitimate recombination events involving gap regions of open circular viral DNA. TPV-L sequences have, like TVCV, been detected in various other *Nicotiana* species including *N. otophora*, *N. sylvestris* and *N. tomentosiformis*, and are also detected in *Datura* and tomato but not in petunia, *Arabidopsis*, or pea (Jakowitsch *et al*., 1999).

III. Petunia vein clearing virus (PVCV)

PVCV is a member of the family *Caulimoviridae* and the coding information is present as one large open reading frame within the viral genome. Data indicate that the entire PVCV genome is present in the *P. hybrida* cv Himmelsroschen genome (Richert-Pöggeler and Shepherd, 1997). The restriction endonuclease pattern and copy number of the PVCV integrant in the white flowered *P. axillaris* ssp. *axillaris* plants resembles more closely those of *P. hybrida* than do those of *P. integrifolia* ssp. *inflata*. The correlation of integrated sequence, stress or physiological change, and detectable virus indicates a similar phenomenon to BSV and TVCV (Harper et al., 2002).

IV. Rice tungro bacilliform virus (RTBV)

There are also two reports by Nagano *et al.,* (2000) and Kunii *et al*., (2004) regarding segments of the rice tungro bacilliform virus (RTBV), in the nuclear genome of rice. Kunii *et al*., (2004) characterized RTBV-like sequences in the Japonica (cv. Nipponbare) genome. These sequences denoted endogenous RTBV-like sequences (ERTBVs), were highly rearranged and dispersed throughout the rice genome. It was found that these sequences were unlikely to have functional potential as a virus, while phylogenetic analysis showed that at least three times integrations of authentic ERTBVs occurred during *Oryza* speciation

1.3.3 DNA transfer between bacteria and plants

Nuclear HGT is rare in multicellular eukaryotes, with most known cases involving bacteria as donors (Garcia-Vallve *et al*., 2000; Rosewich and Kistler, 2000; Screen and St Leger, 2000; Intrieri and Buiatti, 2001; Veronico *et al*., 2001; Watts *et al*., 2001; Wolf and Koonin, 2001; Kondo *et al*., 2002; Zardoya *et al*., 2002; Hall *et al*., 2005). The best characterized example of horizontal DNA transfer between bacteria and plants is the *Agrobacterium* system *(*Figure 1.3). Although plants represent the natural hosts for *Agrobacterium*, this microorganism can also genetically transform a wide range of other eukaryotic species, from yeast (Bundock *et al*., 1995; Piers *et al*., 1996; Sawasaki *et al*., 1998) to fungi (de Groot *et al*., 1998; Gouka *et al*., 1999;

Chen *et al*., 2000; Rho *et al*., 2001) as well as human cells (Kunik *et al*., 2001). Unlike other plant pathogenic bacteria which affect the host-plant physiology by secreting compounds such as toxins or growth regulators, *Agrobacterium tumefaciens* and its related pathogenic species, *A. rhizogenes* and *A. vitis*, directly modify the genetic material of their hosts. This genetic modification results from the transfer and integration into the plant genome of a specific DNA segment termed transferred DNA or T-DNA, from the bacterial Ti (tumor inducing) plasmid (Jouanin, 1984). Plant genetic transformation by *A. tumefaciens* requires the presence of two genetic components located on the bacterial Ti plasmid: (i) T-DNA, the actual genetic element transferred into the plant cell genome; and (ii) the virulence (*vir*) region composed of seven major loci (*virA*, *virB*, *virC*, *virD*, *virE*, *virG* and *virH*), encoding most components of the protein machinery mediating T-DNA transfer. In addition, a set of *A. tumefaciens* chromosomal virulence (*chv*) genes participates in the early stages of the bacterial chemotaxis and attachment to the plant cells (Hooykaas and Beijersbergen, 1994; Sheng and Citovsky, 1996; Zambryski, 1992).

A. rhizogenes is responsible for the formation of adventitious roots known as "hairy roots." The agropine-type strain of *A. rhizogenes* has two distinct T-DNA regions, left transferred DNA (TL-DNA) and right transferred DNA, in its Ri plasmid. Four loci involved in root induction have been identified from an analysis of TL-DNA by insertional mutagenesis. They are designated *rolA*, *rolB*, *rolC*, and *rolD* (White *et al*., 1985). Ng*rol* genes (Ng*rolB*, Ng*rolC*, *NgORF13*, and *NgORF14*) that are similar in sequence to genes in the left transferred DNA (TL-DNA) of *A. rhizogenes* have been found in the genome of untransformed plants of *Nicotiana glauca* (Aoki and Syōno, 1999). It has been suggested that a bacterial infection resulted in transformation of Ng*rol* genes early in the evolution of the genus *Nicotiana.* It has been demonstrated that other bacterial species outside the *Agrobacterium* genus transformed with the Ti-plasmid were also able to transfer T-DNA to facilitate the transfer of foreign DNA to plants (Broothaerts *et al*. 2005). Other examples of bacterium-to-plant nuclear genome HGT include the acquisition of aquaglyceroporins from a eubacterium ~1200 million years ago (Zardoya *et al*., 2002), and of glutathione biosynthesis genes from an alpha-proteobacterium (Copley and Dhillon, 2002), but the mechanisms of transfer remains unclear.

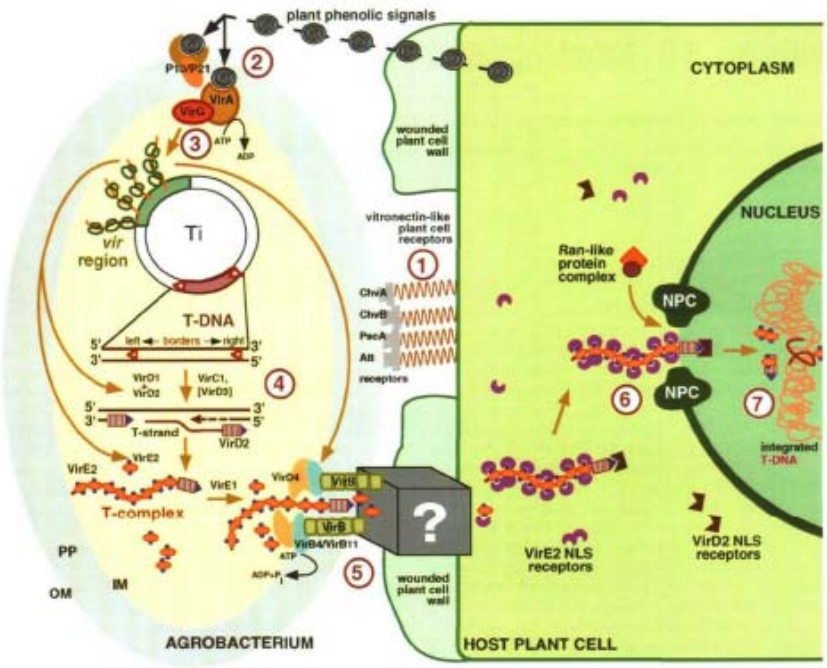


Figure 1.3: A diagrammatic summarization of all major cellular reactions involved in *Agobacterium* T-DNA transport. Steps 1 through 7 indicate sequential processes that occur during Agrobacterium infection. Step 1, binding of Agrobacterium to the host cell surface receptors; step 2, recognition of plant signal molecules by the bacterial *VirA*/*VirG* sensor-transducer system; step 3, activation of the bacterial *vir* genes; step 4, production of the transferable T-strand; step 5, formation of the T-complex and its transport into the host plant cell; step 6, nuclear import of the T-complex; and step 7, T-DNA integration. IM, bacterial inner membrane; NPC, nuclear pore complex; OM, bacterial outer membrane; PP, bacterial periplasm (adapted from: Sheng and Cithosky, 1996).

1.3.4 DNA transfer between fungi and plants

Probably the best example of horizontal transfer to plants from a fungal donor involves the *cox1* mitochondrial intron. During a survey of plant mitochondrial cytochrome oxidase subunit 1 (*cox1*) genes, Vaughn *et al.* (1995) discovered a group I intron in the angiosperm *Peperomia polybotrya*. This was surprising, in so far as only group II introns had previously been found to be associated with plant mitochondria. Phylogenetic analysis revealed different evolutionary histories for the intron and the exon at that locus and clustered the intron together with group I

mitochondrial introns from fungi. It was therefore hypothesized that *P. polybotrya* acquired the intron from a fungal donor. A follow-up study (Adams *et al*., 1998) determined the intron to be present in all *Peperomia* species tested, therefore dating the transfer event(s) before the divergence of the genus. Cho *et al*. (1998) revealed its presence in 48 genera, albeit with an extremely patchy distribution. This is in stark contrast to a nearly universal presence of a cox2 group II mitochondrial intron.

In another example of DNA transfer from fungi to plants, sequences similar to fungal cytoplasmic and mitochondrial virus RNA-dependent RNA polymerase (RdRp) proteins have also been identified in translated protein sequences from the mitochondrial genomes of *Arabidopsis thaliana* and *Vicia faba* (Marienfeld *et al*., 1997). Marienfeld *et al.* (1997) therefore suggested horizontal transfer probably from fungi to those plant species. Plasmids are another genetic element in fungi that are involved in HGT between fungal strains as well as between different species (Taylor, 1986; Collins and Saville 1990; Masel *et al*., 1993; Yang and Griffiths, 1993; Griffiths *et al*., 1994; Arganoza *et al*., 1994; Debets *et al*., 1994; Kempken, 1995; Van der Graag *et al*., 1998). Plasmids have been identified in many fungal species, but are only infrequently encountered in other eukaryotes (Kempken *et al*., 1992; Griffiths, 1995). In yeasts, plasmids are located in the cytoplasm, whereas in filamentous fungi, plasmids are ordinarily associated with mitochondria (Griffiths, 1995). The S1 plasmid of maize appeares to be closely related to linear plasmids of ascomycetes (Kempken *et al*., 1992) and might have been acquired through a horizontal transfer event. However no model exists to explain DNA transfer from fungi to plants.

1.3.5 DNA transfer between plants and plant mitochondria

There is evidence that nuclear transposable elements have moved horizontally on numerous occasions in multicellular eukaryotes. The last couple of years have seen an explosion of studies reporting additional cases of horizontal transfer of genes in plants (Kidwell and Lisch, 2001; Feschotte and Wessler, 2002; Bergthorsson *et al*., 2004; Davis and Wurdack, 2004; Mower *et al*., 2004; Woloszynska *et al*., 2004; Nickrent *et al*., 2004; Davis *et al*., 2005; Schönenberger *et al*., 2005; Diao *et al*., 2006). DNA transfer facilitated by direct plant-to-plant contact through parasitism has emerged as a common mechanism of HGT with several reported HGT events involving parasitic plants as donors (Mower *et al*., 2004; Davis *et al*., 2005) or as

recipients (Davis and Wurdack, 2004; Nickrent *et al*., 2004). However, this mechanism is unlikely to explain all the transfers, as many of the donor and recipient groups do not have a host–parasite relationship. Other possible mechanisms that have been postulated include illegitimate pollination, herbivory, bacterial or viral transfer, uptake of naked DNA in the soil, and fungal pathogens or symbionts (Bergthorsson *et al*., 2003; Won and Renner, 2003; Davis *et al*., 2005). Plant mitochondrial genomes especially experience frequent and evolutionarily widespread horizontal transfer of genes acquired from other eukaryotes in particular from other plants. Table 1.2 gives a summary of mitochondrial genes acquired through horizontal gene transfer. In one case it was shown that *Amborella trichopoda* has acquired, via HGT, partial or full-length copies of 20 of its 31 mitochondrial genes (Bergthorsson *et al*., 2004).

In all but one of the 40 plant-to-plant HGT cases reported thus far, the transferred gene is a mitochondrial gene (encoding a housekeeping respiratory or ribosomal protein), and thus the dominant mode of HGT in plants reported thus far is mitochondrion-to-mitochondrion. The one apparent exception, involving chloroplast *pvs-trnA* in *Phaseolus* (Woloszynska *et al*., 2004), may actually represent mitochondrion-to-mitochondrion transfer too. This is because chloroplast sequences frequently become incorporated into mitochondrial genomes (Unseld *et al*., 1997; Clifton *et al*., 2004; Sugiyama *et al*., 2005), and it is therefore possible that *Phaseolus* acquired this chloroplast sequence via intermediate transfer through the donor's mitochondrial genome.

Although plant mtDNAs usually contain numerous nuclear- and chloroplast-derived sequences (Unseld *et al*., 1997; Clifton *et al*., 2004; Sugiyama *et al*., 2005), there is not yet any good evidence of a plant chloroplast genome containing DNA from other cellular compartments (Rice and Palmer, 2006). Plant mitochondria possess an active DNA uptake system (Koulintchenko *et al*., 2003), though there is no information on a similar system in chloroplast, the ability to transform chloroplasts in intact plant cells indicate that chloroplast might have a similar DNA uptake system (Svab *et al*., 1990). Such an uptake system may be critically important in the incorporation of both foreign and native DNA. Another well-documented difference between the two organelles that may account for their differential susceptibility towards HGT is their tendency to fuse. Plant mitochondria regularly fuse (Arimura *et al*., 2004; Sheahan *et al*., 2005), promoting recombination between parental mitochondrial genomes, whereas chloroplasts very rarely fuse (Kanno *et al*., 1997; Mohapatra *et al*., 1998).

Table 1.2: Published accounts of horizontally acquired genes shown or thought to be located in plant mitochondrial genomes

Citation	Recipient[a]	Donor[b]	Gene
Bergthorsson *et al*. (2003)	*Actinidia*	Monocot	*rps2*
	Amborella	Eudicot	*atp1*
	Betulaceae	Unclear	*rps11*
	Caprifoliaceae	Ranunculales	*rps11*
	Sanguinaria	Monocot	3' *rps11*
Won and Renner (2003)	*Gnetum*	Asterid	*nad1B-C*
Davis and Wurdack (2004)	**Rafflesiaceae**	Vitaceae	*nad1B-C*
Mower *et al*. (2004)	*Plantago*	**Orobanchaceae**	*atp1*
	Plantago	**Convolvulaceae**	*atp1*
Nickrent *et al*. (2004)	**Apodanthaceae**	Fabales	*atp1*
Woloszynska *et al*. (2004)	*Phaseolus*	Angiosperm	cp *pvs-trnA*
Bergthorsson *et al*. (2004)	*Amborella*	Angiosperm[c]	*atp4, atp6, atp8, atp9, ccmB, ccmC, ccmF*$_{N1}$, *cox2* (2x), *cox3, nad1, nad2, nad4, nad5, nad7, rpl16, rps19, sdh4*
		Moss	*cox2, nad2, nad3, nad4, nad5, nad6, nad7*
Schönenberger *et al*. (2005)	*Ternstroemia*	Ericaceae	*atp1*
	Bruinsmia	Cyrillaceae	*atp1*
Davis *et al*. (2005)	*Botrychium*	Santalales	*nad1B-C, matR*

[a] Recipient lineages are indicated by the genus examined, or when multiple related genera were found to share the same foreign gene, the family name. Parasitic plants are in bold.

[b] Donor lineages as best defined by current data. Parasitic plants are in bold.

[c] All but *atp9, nad5, nad7*, and *cox3* are from eudicots, a derived group within angiosperms.

(Table adapted from: Richardson and Palmer, 2007)

1.4 Models for gene transfer and integration

There are no models available for the exact mechanism of transfer and integration of foreign DNA into a new genome. The best studied system is that of *Agrobacterium*-mediated transfer. Though the early events that lead to recognition and transfer are reasonably well characterized (Sheng and Cithosky, 1996) (also see Figure 1.3), the mechanism that leads to DNA integration is still largely unknown. T-DNA integration is the culmination point of the entire process of the *Agrobacterium*-plant cell DNA transfer. T-DNA does not encode enzymatic activities required for integration (Tinland *et al*., 1995; Mysore *et al*., 1998) and plant DNA ligases, must provide these functions (Ziemienowicz *et al*., 2000; Friesner and Britt, 2003). Chilton and Que (2003) provide strong evidence for T-DNA integration into double-stranded breaks created in the plant genome by a transiently expressed rare cutting endonuclease I-*Ceu*I. Nucleotide sequence analysis of the plant DNA/T-DNA junctions indicated that T-DNA integration occurred by a non-homologous end-joining mechanism (Chilton and Que, 2003). Tzfira *et al*. (2003) utilized the double-stranded DNA breaks created by transient expression of another endonuclease, I-*Sce*I, to demonstrate preferential T-DNA integration into these 18 bp long I-*Sce*I recognition sites as determined by sequencing analyses of integration junctions from 620 independent transgenic lines (Tzfira *et al*., 2003). Both studies suggested that T-strands are first converted to double-stranded intermediates and only then are integrated into the plant DNA (Chilton and Que, 2003; Tzfira *et al*., 2003).

Illegitimate recombination has also been proposed as a model for integration (Gheysen *et al*., 1991). Illegitimate recombination is basically a two-step process: DNA ends are first generated and then joined. For T-DNA integration, the ends of the T-DNA would be joined to plant chromosomal breaks. The T-DNA most likely enters the plant as a linear molecule (Zambryski, 1988; Bakkeren *et al*., 1989). Free ends within the plant DNA, which is the other substrate for the reaction, can be generated in a variety of ways, such as errors during replication or repair, or nicks during exposure of single strands in transcription (Roth and Wilson, 1988). Several enzymes involved in these processes, such as topoisomerases I and II, are known to nick or break DNA. There are indications that T-DNA would preferentially integrate in transcriptionally active regions (Koncz *et al*., 1989; Herman *et al*., 1990). This could be explained by the higher likelihood of nicks in these regions, as well as the better accessibility of the target due to unraveling of the nucleosomes of transcribed DNA by comparison with tightly coiled, transcriptionally silent sequences (Patient and Allan, 1989). It has also been hypothesized that

histone H2A plays an important role in illegitimate recombination of T-DNA into the plant genome. Arabidopsis plants with mutants in this gene are recalcitrant to *Agrobacterium* root transformation (Mysore et al., 2000).

The infection cycle of plant viruses, unlike those that infect vertebrates and bacteria, is not known to involve an integration event. In the two groups of DNA viruses that infect plants, the single-stranded DNA *Geminiviridae* replicates via a rolling circle mechanism and the double-stranded DNA *Caulimoviridae*, replicate by reverse transcription, making use of a virally encoded reverse transcriptase (Grierson and Covey, 1988). They are classified as pararetroviruses to distinguish them from true retroviruses, which have RNA genomes. Retrovirus DNA integrates into host chromosomes by means of a virally encoded integrase (Patience *et al*., 1997), pararetroviruses generally lack the gene for this enzyme, and integration is not required for virus replication. With the absence of an integrase enzyme, the integration of viral sequences into the plant genome must take place *via* illegitimate recombination, in cells undergoing active genetic processes (Hull *et al*., 2000).

Fungi have no known infectious viruses analogous to transducing bacteriophage that transport foreign DNA from one individual to another in prokaryotes. However, several experimental systems point to potential ways in which DNA could move between fungi or between fungi and other organisms. Gene transfer through mechanisms similar to DNA transformation appears to take place in culture or in natural settings, as evidenced by transfer of genes and plasmids (Hoffmann *et al*., 1994; Kempken, 1995).

With the ever increasing amount of available sequence and complete genome data such as the rice genome (Goff *et al*. 2002; International Rice Genome Sequencing Project 2005), comparing whole genomes using computational biology offers a new way to search and identify possible instances of DNA transfer and sharing between non-related genomes. The aim of this study was to identify sequences in the rice nuclear genome sharing similarities with non-rice nuclear sequences, originating either from the rice chloroplast or mitochondrial genomes, or from other possible sources such as viruses, bacteria and fungi associated with plants. Since it is possible that ‘accidental’ DNA transfer, in contrast to the specific and directed transfer seen with *Agrobacterium*, might occur as well between rice and its associated microorganisms, an important focus of this study was not to limit the search to possible transferred genes but to asses the total amount of sequence similarities between the different genomes, including

functional and non-functional sequences. The thesis describes the results found when looking at chloroplast and mitochondrial DNA transfer to the rice nuclear genome. It also provides a comparison between the amount and representation of the different organellar genomes in the nuclear genome of rice. It further describes the different searches done in order to identify sequence similarities in the rice nuclear genome and several rice related viruses, tree different bacterial genomes, namely *Bacillus cereus*, *Pseudomonas syringae pv. syringae* and *Xanthomonas oryzae pv. oryzae*, chosen because of their different degrees of association with rice. Lastly it provides a description of the shared sequences between the nuclear genome of rice and that of the rice blast fungi, Magnaporthe grisea.

2

Materials and Methods

2.1 Blast analysis

All homology searches were done against the *Oryza sativa* cv. *japonica* genome database release 16.0 available from TIGR.

Blastn searches were done on a local server against the rice genome (release 16.0 downloaded from the TIGR website) using the following Perl scripts.

Larger genomes were firstly divided into 100 000 bp sections using the splitter script in Perl

Example:

```
splitter -size 100000 -sequence ‘Sequence name’ -outseq ‘Sequence name’.split

 #creates an output file with FASTA records 100000 bases long

 #‘Sequence name’ = name of the sequence file to be divided
```

The subset of sequences was then saved as separate *.fasta* files in a folder named *fastafiles.* These files were searched against the rice genome database on the local server using the following script:

```
use strict;

use Bio::SeqIO;

use Bio::Tools::Run::StandAloneBlast;


my $DB="RICE";          #Which blast database

my $DIR="fastafiles";   #where are the query files
```

```
my @files = `ls $DIR`;  #get a list of all files in $DIR

my $element="";         #init $element;

foreach $element (@files){              #put the next element in @files
into $element

        chomp $element;                 #remove extraneous \n

        my $FILE = "$DIR/$element";     # setup a var with path to file

        my $OUTPUT = "blast.out.".$element;

        my @params = ('program' =>'blastn','database' => $DB,'outfile'
=> $OUTPUT,'_READMETHOD' => 'Blast');

        #blast.out.$element should generate an ouput file for each
        query file so you do not overwrite previous blasts


        my $factory = Bio::Tools::Run::StandAloneBlast->new(@params);

            my $seq_in = Bio::SeqIO->new (-file => $FILE ,'-format' =>
            'fasta');

            my $query = $seq_in->next_seq();


        print "Starting blast on $FILE","\n";     #let the user know
where we are
```

This script resulted in a output file for each sequence file to be searched against the database containing the alignment results (if any). The individual fragments from the different genomes that produced alignments with e-values e-20 or smaller were then realigned with the rice genome database on the *Gramene* genome browser (http://www.gramene.org) to obtain the specific start and end positions on the various rice chromosomes. The results of these alignments were then combined in excel spreadsheets for each genome comparison for further analysis. Mapping the alignment data back onto the karyotype of rice was done using the various facilities provided by *Gramene*.

The specific criteria and sequences used for each analysis are as follows:

2.1.1 Chloroplast ad Mitochondrial Comparison

The *Oryza sativa* cv. *japonica* chloroplast genome (NC_001320) of 134525 bp as well as the mitochondrial genome (BA000029) of 490520 bp were obtained from the *National Center for Biotechnology Information* (NCBI) website (http://www.ncbi.nlm.nih.gov/index.html).

For the chloroplast and mitochondrial genomes the intact sequence were used in homology searches against the rice. Only alignments greater than 100 bp with a homology of 95% and greater were kept for further analysis. Each of these 100bp and greater fragments were then re-aligned with the rice genome using the *Gramene* genome browser. This provided the specific start and end positions of the alignments on the different rice chromosomes. Once again only alignments of 100 bp and greater with and expectation value of e-20 and smaller were used. All repeated sequences within the plastid genomes that aligned to the same positions on the rice genome as well as internal and smaller alignments within larger alignments in the rice chromosomes were removed from the data sets. Where more than one fragment aligned to the same position in the rice nuclear genome, the one with the lower homology or greater e-value were removed.

2.1.2 Viral Comparison

The viral genomes used in the comparison included: Magnaporthe grisea virus 1 (NC_006367); Oryza rufipogon endornavirus (NC_007649); Oryza sativa endornavirus (NC_007647); Rice dwarf virus (Taxonomy id: 10991); Rice gall dwarf virus (Taxonomy id: 10986); Rice black streaked dwarf virus (Taxonomy id: 10990); Rice grassy stunt virus (Taxonomy id: 66266); Rice ragged stunt virus (Taxonomy id: 42275); Rice stripe virus (Taxonomy id: 12331); Rice tungro bacilliform virus (NC_001914); Rice tungro spherical virus (NC_001632); Rice yellow mottle virus (NC_001575); Rice yellow stunt virus (NC_003746) and the Soil-borne cereal mosaic virus (Taxonomy id: 100887); Wheat dwarf virus (NC_003326); Maize streak virus (NC_001346); Sugarcane streak virus (NC_003744); Panicum streak virus (NC_001647); Wheat streak mosaic

virus (NC_001886); Wheat eqlid mosaic virus (NC_009805); Sorghum mosaic virus (NC_004035) and Maize dwarf mosaic virus (NC_003377).

Full length DNA sequences of the viral genomes were submitted to the *Gramene* Blastn server (http://www.gramene.org/multi/blastview) for homology and similarity searches against the rice (*Oryza sativa* cv. *japonica*) genome. Searches were done to identify and include distant homologies with a word size of 9; mismatches scored at -1; gap open penalties at 2 and gap extension penalties at 1. Results were limited to matches with e-values of e^{-5} and smaller and fragments of 100 bp and longer.

2.1.3 Bacterial Comparison

The following bacterial genomes were used in the comparative studies: *Bacillus cereus* ATCC 10987 (NC_003909), *Pseudomonas syringae pv. syringae* B728a (NC_007005) and *Xanthomonas oryzae pv. oryzae* KACC10331 (NC_006834).

The bacterial genomes were divided into 100 Kb fragments that were each aligned with the rice genome using the Blastn algorithm. Only alignments greater than 100 bp were kept for further analysis. Each of these 100 bp and greater fragments were then re-aligned with the rice genome using the *Gramene* genome browser (http://www.gramene.org/multi/blastview). This provided the specific start and end positions of the alignments on the different rice chromosomes. Once again only alignments of 100 bp and greater with and expectation values of e-20 and smaller were used. Further identification and characterization of the sequences were done using the facilities provided by the National centre for biotechnology information (http://www.ncbi.nlm.nih.gov/index.html). All repeated sequences within the bacterial genomes that aligned to the same positions on the rice genome as well as internal and smaller alignments within larger alignments in the rice chromosomes were removed from the data sets.

2.1.4 Fungal Comparison

The *Magnaporthe grisea* genome (Data Version 10/31/2003; Release 2.3) were downloaded from *Magnaporthe* Sequencing Project, Ralph Dean, Fungal Genomics Laboratory at North

Carolina State University (www.fungalgenomics.ncsu.edu), and Centre for Genome Research (www.broad.mit.edu).

The *Magnaporthe* genome was used in the contigs of various lengths in which it was downloaded and aligned against the rice genome on a local server using Perl scripts. Only alignments of 100 bp and greater with and expectation values of e-20 and smaller were used. All repeated sequences within fungal genomes that aligned to the same positions on the rice genome as well as internal and smaller alignments within larger alignments in the rice chromosomes were removed from the data sets. Sequence dendograms were created by multiple alignments using ClustalW and sequence fragments with the best homologies compared to the rice genomic fragments found with Blastn.