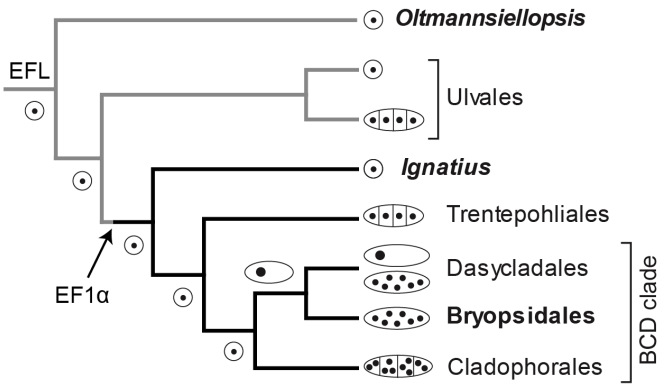
Probing key innovations with next generation sequencing

Key innovations are aspects of organismal phenotypes that promote diversification. They are important adjustments in morphological, cell biological or physiological mechanisms that are essential to the origin and evolutionary success of new groups of organisms.

The evolution of the green seaweeds is characterized by three important innovations. First, the **cytological layout** of the green seaweeds undergoes major transitions. In a paper my colleagues and I recently published in Molecular Biology and Evolution [1], a scenario of cytomorphological evolution is proposed based on a 10-gene phylogeny (see Figure). **Multicellularity** evolved three times independently from single-celled ancestors (Ulvales, Trentepohliales, Cladophorales). Following selection pressures for cell enlargement and macroscopic growth in the ancestor of the BCD clade, two evolutionary lineages emerged. The Cladophorales evolved large cells containing multiple nuclei that each provide transcripts for a well-delimited **cytoplasmic domain**. The common ancestor of Bryopsidales and Dasycladales developed a large tubular cell with cytoplasmic streaming and a greatly enlarged **macronucleus** providing transcripts to the entire cell. This condition is still found in the smaller representatives of the Dasycladales and early life stages of some Bryopsidales. However, a single macronucleus is insufficient to provide transcripts for larger organisms. In the lineages containing large species, a **siphonous cell** architecture with thousands or **millions of nuclei** has evolved. The giant tubular cell of siphonous algae branches and fuses to form complex morphologies with root-, stem- and leaf-like structures. Some of these algae consist of a single cell that is hundreds of kilometers long and forms structures comparable in size to large shrubs on land [2, 3]. Compared to single-celled lineages that contain only a few species (*Oltmannsiellopsis* and *Ignatius*), the macroscopic seaweed lineages have diversified enormously, with ca. 416 species of Ulvales, 378 Cladoporales and 557 Bryopsidales and Dasycladales [4].



A second major cell biological adjustment is situated in the **translational apparatus**. Three sources of evidence support the notion that the translational machinery of green seaweeds has experienced a complete makeover. (1) Recent work shows that **elongation factors**, which deliver tRNAs charged with amino acids to the ribosome, are present in two types [4, 5]. The early-branching lineages of the Chlorophyta all have the elongation factor-like gene (EFL), but the lineage comprising *Ignatius*, Trentepohliales and the BCD clade (see Figure) have elongation factor 1 alpha (EF1α). (2) Recent work in our group also indicates major **shifts of codon usage** in the green algal lineage [5]. In the Dasycladales and Cladophorales, this has even proceeded to the evolution of an **alternative genetic code** [5, 6]. (3) Finally, there is unpublished evidence for **accellerated evolution** of the 18S gene of green seaweeds, especially of the BCD clade [5]. The 18S gene is transcribed to form the small RNA subunit of the ribosome.

The third key innovation is the origin of **C_4 photosynthesis** in the Bryopsidales [6]. The C_4 metabolism nearly eliminates photorespiration and increases fitness in aquatic habitats saturated with O_2. Based on a time-calibrated phylogeny of the Bryopsidales we recently published [7], it seems reasonable to assume that C_4 photosynthesis originated in the ancestor of the "core Halimedineae" during the late Paleozoic glaciation, a period characterized by high atmospheric O_2 and low CO_2. The core Halimedineae have subsequently diversified into tropical habitats, especially in tropical lagoons where O_2 saturation is common. Despite their relatively young age (ca. 250 Ma), the core Halimedineae contain more than half the species of Bryopsidales [4].

It is generally assumed that evolution by natural selection occurs in relatively small steps. However, larger evolutionary steps commonly follow the **origin of new genes by gene duplication** [8]. Because duplicated genes are redundant, one of the copies is free of functional constraint and can evolve a new or more specialized function.

Goal

This project aims to elucidate the genetic correlates of the three key innovations mentioned above by generating expressed sequence tag (EST) libraries of three ulvophycean algae: *Oltmannsiellopsis viridis*, *Ignatius tetrasporus* and *Udotea flabellum* (Bryopsidales). EST libraries represent the pool of genes expressed in the organisms and, following normalization of the cDNA library, both strongly and weakly expressed genes will be well-represented in the EST data. With the availability of multi-species EST datasets, it will become possible to study the evolution of genomic information in an explicit phylogenetic context and gain insight in the three key innovations introduced above.

Global approach and justification

EST data will shortly be available for five of the seven principal lineages in the Figure: Ulvales [9, 10], Dasycladales [11], Trentepohliales, Bryopsidales and Cladophorales (the latter three are being generated in our lab). The three organisms selected for this project were carefully chosen to fill in crucial gaps in our knowledge. When analyzed with the other available data, the three libraries proposed here will permit studying modifications of the pool of coding sequences associated with the three key innovations listed above. In addition to completing the ulvophycean picture, the choice of these three organisms is justified by the following arguments. *Oltmannsiellopsis* is a unicellular organism that provides a reference point for comparison. It diverged from the remainder of the Ulvophyceae before any changes in the cytomorphology or translational apparatus occurred. *Ignatius* is the earliest-branching lineage to have the EF1α gene so it is expected to provide information about the first steps in the makeover of the translation machinery [12]. Furthermore, it is a unicellular sister group of the lineage with deviant cytomorphologies, and as such provides a reference point before any cytomorphological changes had occurred [1]. The choice for *Udotea flabellum* is supported by the fact that this is at present the only species in which C_4 photosynthesis has been shown unambiguously [6].

1. Analysis of the EST data will allow the identification of gene families in association with transitions to multicellularity and the evolution of multinucleate and siphonous cells. Partial information about functional diversification can be gleaned from in-silico analysis [13]. After a selection of apparently diversifying gene families has been made, targeted gene amplification in a broader set of taxa will complete the image of their diversification in association with the cytological diversification. The EST approach offers the advantage that even gene families that would not a priori be expected to be associated with the cytological diversification can be detected.
2. All of the generated data will be used to study the evolution of codon usage and the sequence of steps leading to the evolution of the alternative genetic code of the Dasycladales and Cladophorales. The EST libraries will also provide information on the evolution of ribosomal proteins and eukaryotic release factors, which will advance our knowledge of the makeover of the translation apparatus. Given the important role of ribosomal proteins in organizing the rRNA tertiary structure, it is crucial to assess whether the highly divergent rRNA genes in the giant-celled lineages are correlated with lineage-specific evolutionary trends of ribosomal proteins, such as lateral gene transfer or rate acceleration. The information gained about eukaryotic release factors will enable more profound insights in the evolution of the new genetic code in Dasycladales and Cladophorales.
3. The *Udotea* library will be compared to that of *Bryopsis*, a C_3 species of the Bryopsidales. This comparison will permit identifying duplicated genes of the carbon fixation metabolism. The EST library will be used to design primers and probes to sequence these genes in other bryopsidalean taxa. This will allow pinpointing the gene duplications in evolutionary time and investigate selective regimes on the proteins [14].

I have chosen for next generation sequencing (NGS) technologies because of their favorable cost-benefit ratio. Traditional EST libraries involve introducing the cDNA in vectors, transforming bacterial cultures and sequencing the resulting clones with Sanger technology. NGS is based on innovations in PCR technologies, which requires significantly less labor and is consequently much cheaper. The greatest advantage of NGS technologies is that, for a given budget, it yields several orders of magnitude more data than Sanger methods,

thereby providing much higher coverage of the coding sequence information in the genome. We chose for Solexa Illumina sequencing to carry out this project. While the reads of this technology are fairly short, the number of reads is incredibly high, resulting in good contig assembly for transcriptome sequencing. Extrapolating from knowledge about the model species *Arabidopsis*, I anticipate that the proposed four lanes of Solexa sequencing will yield a nearly complete coverage of the transcriptome with sufficient sequencing depth [15].

Summary of methods

Strains of *Oltmannsiellopsis viridis* (NIES-360) and *Ignatius tetrasporus* (UTEX-B2012) will be grown in culture flasks. *Udotea flabellum* will be frozen in liquid nitrogen in the field. RNA will be extracted following protocols developed and tested in our lab.

Construction and normalization of a cDNA library and next-generation sequencing will be outsourced to Cofactor Genomics, our regular NGS partner [16]. Each normalized cDNA library will be sheared to an appropriate fragment length and sequenced on four lanes of a Solexa Illumina Genome Analyzer 2 [17]. Cofactor Genomics also provides the primary bioinformatics for contig assembly using Velvet [18].

The subsequent analysis steps differ somewhat depending on the particular goal, but a few commonalities can be identified. After the generated data has been combined with previously sequenced EST libraries, Markov clustering or superparamagnetic clustering will be used to delimit gene families [19, 20]. Subsequently, the number of member genes belonging to any given gene family will be will be counted for each species and the gene diversity compared among species. This will give indications as to which gene families have diversified. Subsequently, phylogenies of the diversifying gene families will be made to pinpoint where gene duplications (or possibly whole genome) duplications occurred [21]. After interesting gene families have been identified, they will be functionally characterized with in silico methods [22-24]. Finally, the selective regimes acting on the different gene copies during gene family diversification will be examined using mathematical models of coding sequence evolution [25, 26].

Project-specific elements include analysis of codon usage and detailed studies of certain genes or gene families of interest. The study of changes in the translational machinery encompasses detailed analyses of synonymous codon usage, which can be quantified for each species with the SCUO measure [27]. The evolution of codon usage will subsequently be modeled in a phylogenetic context to pinpoint where and when transitions occurred [5, 26]. Genes and gene families targeted for in depth studies are PEPC, PEPCK, CA and other genes involved in carbon fixation, eukaryotic release factors that can be anticipated to play a crucial role in the evolution of alternative genetic codes [28], and families of ribosomal proteins that can be expected to show erratic patterns of evolution in association with the makeover of the translational apparatus.

References

[1] Cocquyt E, et al. 2010. Evolution and cytological diversification of the green seaweeds (Ulvophyceae). *Mol Biol Evol* 27:in press. **[2]** Vroom PS, Smith CM. 2003. Life without cells. *Biologist* 50:222-226. **[3]** Littler MM, et al. 2005. Extraordinary mound building Avrainvillea (Chlorophyta): the largest tropical marine plants. *Coral Reefs* 24:555-555. **[4]** Guiry MD, Guiry GM. 2010. AlgaeBase. World-wide electronic publication. [http://www.algaebase.org] **[5]** Cocquyt E. 2009. Phylogeny and molecular evolution of green algae. *PhD thesis.* Ghent University, Phycology Research Group. **[6]** Reiskind JB, Bowes G. 1991. The role of phosphoenolpyruvate carboxykinase in a marine macroalga with C4-like photosynthetic characteristics. *Proc Natl Acad Sci USA* 88:2883-2887. **[7]** Verbruggen H, et al. 2009. A multi-locus time-calibrated phylogeny of the siphonous green algae. *Mol Phylogenet Evol* 50:642-653. **[8]** Carroll SB. 2005. Evolution at two levels: On genes and form. *PLoS Biol* 3:1159-1166. **[9]** Stanley MS, et al. 2005. Analysis of expressed sequence tags from the green alga *Ulva linza* (Chlorophyta). *J Phycol* 41:1219-1226. **[10]** Niu J, et al. 2010. Analysis of expressed sequence tags from the *Ulva prolifera* (Chlorophyta). *Chin J Oceanol Limnol* 28:26-36. **[11]** Henry I, et al. 2004. Comparison of ESTs from juvenile and adult phases of the giant unicellular green alga *Acetabularia acetabulum*. *BMC Plant Biol* 4:3. **[12]** Cocquyt E, et al. 2009. Gain and loss of elongation factor genes in green algae. *BMC Evol Biol* 9:39. **[13]** Raes J, Van de Peer Y. 2003. Gene duplications, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl Bioinf* 2:92-101. **[14]** Besnard G, et al. 2009. Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): Multiple appearances and genetic convergence. *Mol Biol Evol* 26:1909-1919. **[15]** Wall PK, et al. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 10:347. **[16]** Cofactor Genomics: http://www.cofactorgenomics.com **[17]** Solexa Illumina GA2: http://www.illumina.com/systems/genome_analyzer.ilmn **[18]** Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821-829. **[19]** Enright AJ, et al. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575-1584. **[20]** Tetko IV, et al. 2005. Super paramagnetic clustering of protein sequences. *BMC Bioinf* 6:82. **[21]** Van de Peer Y, et al. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10:725-732. **[22]** Nagaraj SH, et al. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics* 8:6-21. **[23]** Gibbons JG, et al. 2009. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol Biol Evol* 26:2731-2744. **[24]** Engelhardt BE, et al. 2005. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol* 1:e45. **[25]** Yang ZH, et al. 2000. Codon-

substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431-449. **[26]** Yang ZH, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* 25:568-579. **[27]** Angellotti MC, et al. 2007. CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res* 35:W132-W136. **[28]** Keeling PJ, Leander BS. 2003. Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*. *J Mol Biol* 326:1337-1349.