

The *Physcomitrella* genome project: Workshop discussion at Moss 2004

Brent Mishler: Overview

1. The US Department of Energy is to undertake sequencing of the *Physcomitrella* genome, under the auspices of its Joint Genome Institute (JGI): a part of the “Community Sequencing Programme”. This follows the successful submission of a proposal by Brent Mishler (University of California, Berkeley) and Ralph Quatrano (Washington University at St. Louis) as joint Principal Investigators, supported by collaborators Ralf Reski (University of Freiburg, Germany), Mitsuyasu Hasebe (National Institute for Basic Biology, Japan), David Cove and Andrew Cuming (Leeds University, UK).
2. The JGI project will undertake shotgun sequencing of phosmid-cloned genomic DNA to 6x coverage, assembly of the sequence into contigs and provide “some help” with finishing.
3. The community must seek and obtain funding for the actual finishing of the sequence, annotation and informatics (including the establishment of accessible databases), central resource collection and dissemination (a culture collection) and functional studies for comparative genomics.
4. The programme will enable phylogenetic studies, through provision of new genes for systematic analysis and the identification of new genomic characters. It will drive comparative genomic studies, with phylogenetic relationships informing further bioinformatic analysis. It will enable functional genomic approaches on a truly genome-wide scale.
5. Particular problems that will be addressed relate to the diversity of green plants. Although probably over 500,000 species exist, we know little of the mechanisms that underly this diversity. Questions that the *Physcomitrella* genome will allow us to address include the origin of multicellularity, the establishment of reproductive strategies and the nature of environmental stress tolerances, the evolution of gene function, population genetics, questions relating evolution and development, the co-evolution of nuclear and organellar genomes, physiological ecology through gene-to-phenotype-to phylogeny approaches.
6. *Physcomitrella* will be the first haploid plant genome to be sequenced. This will make annotation of the sequence easier and may provide additional insight into how haploid genomes maintain their stability, and to the capacity of *Physcomitrella* for high rates of homologous recombination.

How the community can best interact with the genome project.

Major research projects of this type provide substantial leverage for the acquisition of further research funding. The recommended model to follow is that of the “Deep Green” consortium which set out to establish the evolutionary relationships within the extant flora. This programme established a series of meetings and extensive web

based data-sharing that enabled the construction of the current phylogeny. Crucially, it gave rise to a variety of spin-off programmes. These included

- The Green Tree of Life
- Deep Gene
- Deepest Gene
- Deep Time (the fossil record)
- The Green Plant BAC library collection
- The Floral Genome Project
- Chloroplast genomics
- The JGI *Physcomitrella* and *Selaginella* programmes

The success of these “offspring” resulted from the members and supporters of the original consortium **participating, collaborating and supporting** each other’s proposals, rather than competing for limited resources.

To follow this model we should similarly establish the *Physcomitrella* genome programme as a loose consortium with no formal director, but one in which all the members fully communicate their research and support each others funding proposals with strongly worded letters of support and positive reviews (subject to scientific quality!)

Ralph Quatrano: Update from DoE

It is necessary to establish a “DNA pipeline” in which materials will be provided to the sequencing labs, who will in turn generate the sequence data. At present, the specifics are not clear as to the resources required and the form of the data output. In the immediate future, BM, RSQ and DJC will visit the DoE laboratories to establish the requirements and clarify the nature of the data acquisition and output processes.

The first step will be to provide bulk preparations of pure nuclear DNA. This will undergo quality control by DoE. Once this has been done, the timeline for the process will be clearer.

An important question will be the provenance of the DNA to be sequenced. It should be from a single spore-derived culture.

The Genome Center at Washington University has had extensive experience in genome analysis, particularly in the Human Genome Project, and has expressed a readiness to interact with JGI. Other useful benefit could be gained through closer interactions with other plant genome consortia.

It will be necessary to obtain substantial additional funding. In particular this will be necessary for the “finishing”, assembly and annotation processes. We must target the appropriate sponsors, and mutually support one anothers’ applications.

In particular, we must address the question brought up by all funding bodies:

“How will the sequence be used, who will use it and how does it benefit our particular interest group?”

A further resource available to the community is a *Physcomitrella* gene chip. This is based on the publicly available EST sequence data, and comprises 22,000 60-mer oligonucleotides. The design of the chip was supported by the Leeds-Wash. U. PEP programme, in collaboration with “Mogen”, a company set up in St. Louis.

The chips are manufactured by Agilent Technologies and are printed using ink-jet technology. To order chips, submit a request to Mogen (*i.e.* Ralph Quatrano). Currently a 30% discount is being offered to academic users (to 15th April), so the basic chip price is \$490 each. Mogen additionally offer a hybridisation and analysis service as an Agilent approved service which would bring the total cost for the full package to approximately \$1,000: alternatively you can undertake your own analysis.

Mitsuyasu Hasebe: The Japanese dimension

Further funding has been obtained to sustain *Physcomitrella* genomic research until 2009. This is derived from NIBB and MEXT (Ministry of Education, Sports, Culture & Technology). This will be delivered through a programme led by MH and **Tomoaki Nishiyama**.

In the first 5-year period, we have seen the establishment of substantial resources:

2000-2005: Full-length cDNA libraries were constructed and an EST database established.

By 2005 there will be 5’ - and 3’ - sequences determined from all the extant libraries, that include protonemata (including auxin and cytokinin treatments), and gametophore tissue, and also from new libraries representing regenerating protoplasts, early (premeiotic) and late (postmeiotic) sporophyte development, induced and differentiating gametangia

2005-2006: Full-length sequences from the protonemal, regenerating protoplast, sporophyte and gametangia cDNA libraries will be completed.

By end 2004: BAC and phosmid libraries will be constructed to 20x genome coverage

By end 2005: BAC end-sequencing of 100,000 clones (corresponding to 100Mbp) will be completed, permitting assembly with the sequence obtained from the JGI programme (subject to the speed of data release)

By end 2006, annotation based on placing cDNA sequences on a basic scaffold will be achieved.

Post-genomic developments

The period 1998-2000 saw the construction of a collection of tagged gene-trap and enhancer-trap lines of *Physcomitrella*, for functional studies. This will be supplemented by:

2006: An oligonucleotide-based microarray derived from all annotated genes

2006-2009: Generation of *ca.* 2000 GFP/GUS fusion lines based on reporter insertion at the 3'-ends of selected genes (transcription factors, receptors, kinases, signal transduction intermediates *etc.*).

Stefan Rensing: The Bioinformatic Challenge

At Freiburg, assembly and annotation of EST sequences has been based on the substantial resources available: both public (NIBB, Leeds) and proprietary (BASF). Taken together, this places *Physcomitrella* sixth in a "league table" of plant species. (Above *Arabidopsis*, below wheat, rice, maize, barley and soybean).

Moss-specific UTR repeats have been identified and masked for clustering of the public EST dataset. Likewise, a HMM/SVM (hidden Markov model/ support vector machine)- based prediction of splice sites in *Physcomitrella* has been established. These resources are publicly available via www.cosmoss.org

These resources are already well-used by the community (more than 3,500 accessions in 2004).

Assembly and annotation of the genome requires the following inputs:

- (i) The whole genome shotgun sequence
- (ii) BAC-end skim sequences
- (iii) Assembly of initial long contigs.

This is only the beginning.

- (iv) Clustered EST and full-length cDNA sequences will facilitate annotation and Gene I.D. assignment.
- (v) The genome will contain many repeats: knowledge of their nature is essential
- (vi) A genetic map integrated with the physical map is necessary for the location of genes onto chromosomes
- (vii) Splice-site prediction and gene prediction software must be trained using *Physcomitrella* information. In establishing an annotation pipeline, the data from the clustered EST sets and full-length cDNA sequences will enable *Physcomitrella* specific algorithms for splice site and UTR prediction to be rapidly developed.
- (viii) BAC-end sequences and a genetic linkage map will be required for the assembly of large contigs. The generation of a good integrated genetic and physical map allows the placement of uncertain regions. Development of FISH technology, though challenging, will be essential for this process.

The ultimate goal will be to present the genome through an easy-to-use graphical interface readily accessible to non-specialist bioinformatics user.

Stephane Rombauts: Experience from other genome projects

At Gent, the EUGENE plant systems biology group has been involved in a number of related plant bioinformatics projects. These include:

Arabidopsis

Populus balsamifera

Medicago trunculata

Ostreococcus (a green alga)

And the fungus, *Laccaria bicolor*

Even in genomes that have been well characterised, there remain discrepancies. For example the analyses of the *Arabidopsis* genome by TIGR and by EUGENE - although broadly in agreement - provide differing estimates of gene number (TIGR predicts 26,207, EUGENE predicts 27,979). Work is still ongoing to provide a correct annotation, and the different prediction methodologies used in these two centres provide a valuable complementary approach. The same is true in the annotation of the poplar genome, which is being undertaken as an international consortium that includes a substantial JGI contribution. Here, three participants, ORNL (Oak Ridge National Laboratory), JGI and Ghent have found that by using different but complementary software to undertake gene prediction they have been able to correct and refine each others predictions. In this case, final annotation used a rule-system developed by JGI to elect the “best” annotation.

The poplar programme has many lessons for the *Physcomitrella* community, since it has also been carried through with substantial input from JGI.

In this programme, JGI undertook the DNA sequencing and assembled the sequence into contigs. Meanwhile, ORNL collected linkage data and assembled linkage maps. Activities in the programme were coordinated by frequent conference calls which aided in the assembly of the genome.

Free availability of the data from JGI via an FTP site which permitted other annotators to have access and deposit rights ensured widespread cooperation in assembly and annotation.

This programme used extensive data sets from the very beginning of the project, which, from a bioinformatics point of view, was very helpful in training the software. This largely used *Arabidopsis* models, together with substantial EST data to map all the ESTs on the genome data sets.

The bottom line is that EST data is VERY important in developing trained software that can recognise species-specific intron/exon and UTR sequences.

The *Medicago* project is another example of a well organised collaboration, between INRA-Toulouse, Ghent and MIPS.

Unlike the shotgun-sequencing approaches used in other projects, this has been based on the sequencing of BAC clones. This provides data of much higher overall quality, but the data acquisition is slower.

In this programme, gene prediction programs and BLAST analysis was used, in conjunction with training sets for coding sequence recognition. The EUGENE consortium has a lot of experience in developing trained software for gene recognition.

***Ostreococcus*: an unusual genome**

This is a green alga with a very small, compact genome. It exhibits several unusual features:

- Most genes comprise a single exon
- Those that do not fall into one of two classes: multiple small-intron (36-70bp) genes with unusual splice junctions, all located on a single chromosome, and genes with fewer but larger introns (80-500bp) that have canonical splice junctions.
- Close linkage of related genes (*e.g.* those associated with nitrogen fixation)
- Overlapping protein coding sequences of adjacent genes

Prospects for *Physcomitrella*

The Ghent group has a wealth of experience in gene recognition and annotation software development, and in addition to coding sequence recognition has experience in gene comparison, transposable element recognition, microRNA prediction, functional analysis and promoter prediction.

This group can host a suitably qualified individual to learn and use all of these tools and apply them to the *Physcomitrella* data sets as they become available. It is estimated that the necessary training period would be for two months at a minimum.

Ghent and Freiburg are already co-operating in this, and a person from Freiburg (**Daniel Lang**) is prepared to go.

Heinz Himmelbauer: Physical mapping of the genome

This group at MPI-Berlin has extensive experience in the generation of physical maps of genomes, derived from collaboration in the mouse, medaka ("rice fish") and sugar beet genome programmes.

The mouse genome map was developed using YAC and BAC clones which were ordered using intervening repeat sequence probes to identify YACs. Additionally, 2000 genetically mapped markers were integrated into the physical map by hybridisation.

Oligonucleotide hybridisation probes provide powerful tools for physical mapping. These can be obtained from any genetically mapped cloned marker, EST or cDNA sequence. For the medaka genome, pools of 35-mer oligos were hybridised to a 36,864 clone BAC array. By using pooled probes in a 3-D array format, data can be

acquired very rapidly. BAC resources from three different fish strains were utilised, and a first-generation-map that encompassed the medaka genome in ~ 900 contigs was constructed. Further effort is currently put into developing an advanced map with higher marker content and lower number of contigs. However, the map has already been of great use to accelerate the map-based cloning of genes that had given interesting phenotypes in medaka ENU-mutagenesis screens. In addition, BAC-based sequencing of medaka chromosome 22 is carried out in Shimizu's group at Keio University.

The sugar beet programme is currently at the stage of developing a “phase 1 map”. 10,000 35-mer probes derived from BAC-end sequences, ESTs and genetic markers are being hybridised to *ca.* 37,000 BAC clones.

A similar approach could be used for *Physcomitrella*:

A Phase-1 map could be derived using 6,000 BAC-end probes and 1000 EST-based probes. This should generate approximately 700 contigs of *ca.* 700kb each. Additionally, some 1500 genetic markers would be required.

The next stage would be to develop a smaller number of long contigs: 1400 probes derived from the phase-1 contig ends would enable the assembly of 350 large contigs of 1.4Mbp

The Genetic linkage map

The *Physcomitrella* genome project is unusual in that no genetic linkage map has yet been derived for this species. However, we learnt that efforts are now in progress to rectify this. **Mark von Stackelberg** (Freiburg) described how the collection and analysis of a genetically diverse collection of *Physcomitrella* accessions has been initiated. In addition to the original “Gransden” isolate, *Physcomitrella* ecotypes have now been obtained from Europe, Australia, Africa, Japan and the USA. While some of these accessions are classified as *P. patens* ssp *patens*, the collection also includes examples classified as subspecies *californica*, *magdalenae* and *readeri*. Initial analyses of genome size by flow cytometry indicates all accessions to have approximately the same genome size, although the *readeri* accessions seem to have a slight but significant larger DNA content. Some variation in the subspecies is apparent following amplification of rDNA ITS sequences – the *readeri* and *californica* subspecies ITS sequences are some 30bp longer than those of the *patens* accessions. Sequencing of ITS amplicons revealed a lot of variant sequences within individual accessions, that may indicate the occurrence of recent hybridisation events.

Following a bioinformatics search for simple sequence repeats (SSRs or “microsatellites”) in the EST database 3729 candidate loci were identified. Of these, 3063 can have good primers designed using primer design software. To date, primer pairs for 71 loci have been used to examine polymorphism between 18 ecotypes and two outgroups (*Funaria* and *Physcomitrium*): high levels of polymorphism have been identified within the collection, with the accessions falling into two distinct phylogenetic groupings: the subspecies *patens* form one group, and the *readeri/californica* group forming the second. 90% of the loci examined were variable between these two groups.

To date, the most promising ecotype within the *ssp. patens* group is a French accession from Haute Saône that exhibits 50% polymorphism with respect to the Gransden strain.

All *Physcomitrella* accessions are archived in Freiburg, documented in www.cosmoss.org and will be distributed to interested scientists.

In a complementary approach, **Andy Cuming** (Leeds) will be generating AFLP markers for linkage mapping. These markers have been chosen because (i) they can be rapidly generated in large numbers (ii) they exhibit high frequencies of polymorphisms, compared with other markers (10-15 polymorphic loci can be identified with any primer pair combination) (iii) they can be rapidly analysed using sequencing gel technology (iv) unlike the SSR markers, AFLP markers are independent of previously determined sequences, and it is anticipated that the bulk of such polymorphic markers will fall in non-coding regions of the genome (v) since these markers are identifiable as DNA fragments on gels, they can readily be cloned for the generation of hybridisation probes in linking to the physical map.

That AFLP markers can be used for the rapid generation of a linkage map was demonstrated in a presentation by **Stuart McDaniel** (Duke University) who has produced a linkage map of *Ceratodon purpureus* for the identification of QTL loci over a period of a few months.

Crosses between the Gransden strain and the other ecotypes, (and between combinations of the ecotypes) have been set up in Freiburg, and by **David Cove** at Leeds, and sporophytes are developing. However, at this point it has not yet been determined whether these derive from self- or cross-fertilization events. This will be clearer over the next month, as spores are recovered and germinated for genotype analysis.

The haploid nature of the dominant gametophyte will substantially assist in the production of mapping lines: because meiosis (and recombination) occurs following the production of the sporophyte, to generate haploid spores, every F1 progeny plant represents a “recombinant haploid” line in which recombination junctions are fixed: thus one generation is sufficient to generate the equivalent of the “recombinant inbred” lines used for the mapping of diploid species, which characteristically require 6 or 7 generations of selfing to fix their genotypes to homozygosity.

An important aim of the mapping project will be to bulk up the RH lines obtained to be made available as a resource for the community.

Workshop discussion

1. Which genotype should be sequenced?

It was generally agreed that the genome sequence should be based on a single genotype, and that this should be the Gransden strain, first isolated by Harold Whitehouse, this being the principal lab. strain disseminated worldwide by David Cove. It was also agreed that the strain should have recently been isolated from a single spore, rather than having been maintained in asexual culture for many years

(which leads to loss of fertility and (probably) the acquisition of somatic mutation).

The current spore-derived culture in use at Leeds was thus accepted as the basis for all future analysis. This strain is to be designated the “**Gransden 2004**” isolate. It can be obtained from Leeds by contacting **Yasuko Kamisugi** (bmbyk@leeds.ac.uk).

Leeds can also provide transgenically marked versions of this strain, containing a single-copy selectable marker gene at a known locus. These versions should be ideal for crossing, since the segregation of the transgenic marker will identify sporophytes derived from a cross, rather than a self-fertilization event. Currently available strains are marked in two different loci with an *nptii* cassette. Strains carrying a hygromycin resistance marker are being constructed.

(Some of the crosses in progress have used strains genetically marked with auxotrophic or chlorophyll-deficient mutations. However, these parental strains were generated by NTG mutagenesis, and are likely to carry a large number of other, uncharacterised mutations).

2. Consortium and communication

For future funding, it will be of assistance for applicants to be recognised as members of the *Physcomitrella* genome consortium. Following the “Deep Green” precedent, it was felt that the consortium should be a loose association, and groups were invited to identify themselves with the consortium. A first step should be the setting up of a web page for the grouping.

Breaking news: Ralph Quatrano has secured the URL domain name “mossgenome.org”.

Interested scientists who officially need the label as being member of the Mossgenome Consortium (*e.g.* to promote their own grant applications) should identify themselves to Brent Mishler (bmishler@socrates.berkeley.edu).

3. Funding

Additional funding will be necessary to maximise the benefits of the sequencing programme, and in turn, the investment in the sequence should be used to leverage support from individual national bodies.

Ralf Reski said that following BASF’s substantial investment in EST resources it was unlikely that the company would wish to support *Physcomitrella* genome sequencing without there being IP options for this company. This would conflict with the requirements of the JGI for direct public accessibility. Consequently, Ralf has submitted a proposal to public German agencies but has no decision on it as yet.

Svetlana Zoriniantz indicated that the Greilhuber group at the University of Vienna would be applying for Austrian funding to support molecular cytogenetic analysis of *Physcomitrella*, and **Jon Hughes** (University of Giessen) plans to

apply to the DFG in the near future for funding for a yeast-2-hybrid based analysis of phytochrome interacting factors, particularly in the cytoplasm. Parties interested in developing a coordinated approach to protein-protein interactions - perhaps also using B2H, split YFP, surface plasmon resonance or other approaches - were invited to contact him.

It would be highly desirable to set up a culture collection and distribution centre for all programme-related materials. Freiburg has led the way in developing cryogenic storage of moss tissue, and might be a suitable focus. Ralf Reski offered that this facility can be used as stock centre for all mosses and this was accepted by the workshop-participants.

Mitsuyasu Hasebe announced, that in Japan the liverwort *Marchantia* will be sequenced; after some discussion about the possible establishment of a joint stock centre, it was generally accepted that a stock centre for liverworts should be established by different resources, e.g. by those, who participate in *Marchantia* sequencing. Nevertheless, we should seek to establish close links with the *Marchantia* programme as well as with any other non-vascular plant genome programmes (*Selaginella* is also to be sequenced by JGI). It was suggested that although previous applications to the Human Frontiers Science Programme had not been successful, this might be an opportune time to investigate further applications to this organisation.

4. Moss 2005

Next year, the International Botanical Congress meets in Vienna (17th-23rd July), and **Ralf Reski** has been invited to organise a symposium on moss for this meeting. It was agreed that the Moss 2005 meeting should be organised as an adjunct to this. **David Cove** suggested that the Mendel Museum in Brno might be a suitable site for this, and would investigate whether the facilities at that site would be appropriate. Ideally, Moss 2005 should be AFTER the IBC 2005, as mentioned by US colleagues, as any other time might conflict with the ASPP meeting.