The Populus Genome Science Plan 2004-2009:

From Draft Sequence to a Catalogue of All Genes through the Advancement of Genomics Tools

The *Populus* Genome Science Plan 2004-2009: From Draft Sequence to a Catalogue of All Genes through the Advancement of Genomics Tools

Table of Contents

Genome Annotation and Assembly Steering Committee

The purpose of the Genome Assembly & Annotation Steering Committee is to assist in the coordination and integration of the DOE Joint Genome Institute sequencing of the poplar genome with structural and functional genomics research taking place in *Populus* and other organisms worldwide.

Brian Ellis, University of British Columbia, Canada

Stefan Janssen, Swedish University of Agricultural Sciences, Sweden

Dan Rokhsar, DOE Joint Genome Institute, USA

Jerry Tuskan, Oak Ridge National Laboratory, USA

Acknowledgments

Sequencing and Annotation Institutional Sponsors and Partners:

DOE Office of Science Biological and Environmental Research Program:
Contact: Marvin Frazier, marvin.frazier@science.doe.gov
The International Populus Genome Consortium:
Contact: Gerald Tuskan, 865-576-8141; tuskanga@ornl.gov
DOE Joint Genome Institute:
Contact: David Gilbert, 925-296-5643; gilbert21@llnl.gov
Genome Canada:
Contact: Anie Perrault, 613-751-4460, ext. 13; aperrault@GENOMECANADA.CA
Genome British Columbia:
Contact: Linda Bartz, 604-637-4373; lbartz@genomebc.ca
Umeå Plant Science Centre:
Contact: Stefan Jansson, +46-90-7865354; stefan.jansson@plantphys.umu.se
Stanford Human Genome Center:
Contact: Ruthann Richter, 650-725-3900; richter1@stanford.edu
Department of Plant Systems Biology and INRA-associated laboratory at Ghent University:
Contact: Yves Van de Peer, +32 (0)9-331-3807; yves.vandepeer@psb.ugent.be
Contact: Pierre Rouzé, +32 476 638 304; pierre.rouze@psb.ugent.be

Genome Annotation and Assembly Participants:
Jörg Bohlmann, Department of Botany, University of British Columbia
Jarrad Chapmann, DOE's Joint Genome Institute
Stephen DiFazio, Plant Genome Group, Environmental Sciences Division, Oak Ridge National Laboratory
Carl Douglas, Department of Botany, University of British Columbia
Brian Ellis, University of British Columbia, Canada
David Gilbert, DOE's Joint Genome Institute
Igor Grigoriev, DOE's Joint Genome Institute
Jane Grimwood, Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine
Stefan Janssen, Swedish University of Agricultural Sciences, Sweden
Jan Karlsson, Umeå Plant Sciences
Miriam Land, Computational Biology Institute, Life Sciences Division, Oak Ridge National Laboratory
Frank Larimer, Computational Biology Institute, Life Sciences Division, Oak Ridge National Laboratory
Philip F. Locascio, Computational Biology Institute, Life Sciences Division, Oak Ridge National Laboratory
Nik Putnam, DOE's Joint Genome Institute
Steven Ralph, BC Cancer Research Centre
Dan Rokhsar, DOE Joint Genome Institute, USA
Stephane Rombauts, Plant Systems Biology, Ghent University
Pierre Rouzé, Plant Systems Biology, Ghent University

Asaf Salamov, LBL DOE's Joint Genome Institute
Jacquie Schein, Genome Sciences Centre, BC Cancer Research Centre
Jeremy Schmutz Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine
Astrid Terry, DOE's Joint Genome Institute
Jerry Tuskan, Oak Ridge National Laboratory, USA
Yves Van de Peer, Plant Systems Biology, Ghent University

Cover Design:
Mark Schuster, University of Tennessee

Cover photo credits:

Greenhouse-grown 'Nisqually-1'	Mark Schuster, University of Tennessee
Hybrid poplar MCAT scan image	Joanne Ledford, Oak Ridge National Laboratory
Populus FISH image	M. Nurul Islam-Faridi, Texas A&M University
Populus tissue culture	Max Cheng, University of Tennessee

Contributors to the Science Plan

Genetic Resources

John Davis	jmdavis@ufl.edu	University of Florida
Jörg Bohlmann	bohlmann@interchange.ubc.ca	University of British Columbia
Stefan Jansson	Stefan.jansson@plantphys.umu.se	Umeå Plant Science Centre
Les Pearson	lxpears@arborgen.com	ArborGen
Barb Thomas	bthomas@ualberta.ca	University of Alberta
Chung-Jui Tsai	chtsai@mtu.edu	Michigan Tech University
Marc Villar	villar@orleans.inra.fr	INRA - Orleans

High-Throughput Phenotyping

Mark Davis	mark_davis@NREL.gov	National Renewable Energy Lab
Bob Kodrzycki	rjkodrz@arborgen.com	ArborGen
Laigeng Li	Laigeng_Li@ncsu.edu	NCSU
Gary Peter	gfpeter@ufl.edu	University of Florida
Simon Potter	Simon.Potter@csiro.au	CSIRO
Lacey Samuels	lsamuels@interchange.ubc.ca	University of British Columbia
Steve Strauss	steve.strauss@oregonstate.edu	Oregon State University
Björn Sundberg	Bjorn.Sundberg@genfys.slu.se	Umeå Plant Science Centre
Gail Taylor	g.taylor@soton.ac.uk	University of Southampton
Stan Wullschleger	wullschlegsd@ornl.gov	Oak Ridge National Laboratory

Physical Mapping & Marker Development

Patricia Faivre-Rampant	Patricia.Faivre-Rampant@evry.inra.fr	INRA - URGV
Rishi Bhalerao	rishi.bhalerao@genfys.slu.se	Umeå Plant Science Centre
John Carlson	jec16@psu.edu	Penn State University
Carl Douglas	cdouglas@interchange.ubc.ca	University of British Columbia
Berthold Heinze	Berthold.Heinze@fbva.bmlf.gv.at	Fed. Res. Centre for Forests
Maurizio Sabatti	sabatti@unitus.it	Univ. degli Studi della Tuscia
Jerry Tuskan	tuskanga@ornl.gov	Oak Ridge National Laboratory

Tissue Culture & Transformation

Rick Meilan	rmeilan@purdue.edu	Purdue University
Malcolm Campbell	malcolm.campbell@plants.ox.ac.uk	Oxford University
Shujun Chang	sxchang@arborgen.com	ArborGen
Takashi Hibino	takashi-hibino@ojipaper.co.jp	Oji Paper Co.
Ove Nilsson	Ove.Nilsson@genfys.slu.se	Umeå Plant Science Centre
Gilles Pilate	pilate@orleans.inra.fr	INRA – Orleans

Metabolic Characterization & Metabolomics

Tim Tschaplinski	tschaplinstj@ornl.gov	Oak Ridge National Laboratory
Thomas Moritz	thomas.moritz@genfys.slu.se	Umeå Plant Science Centre
Andrea Polle	apolle@gwdg.de	University of Gottingen
Scott Harding	sahardin@mtu.edu	Michigan Tech University
Janice Cooke	janice.cooke@rsvs.ulaval.ca	Université Laval
Reinhard Jetter	jetter@interchange.ubc.ca	University of British Columbia

Protein Characterization & Proteomics

Christophe Plomion	plomion@pierroton.inra.fr	INRA - Bordeaux
Fred Sterky	fredrik.sterky@biotech.kth.se	Umeå Plant Science Centre
Cetin Yuceer	mcyl@ra.msstate.edu	Mississippi State University
Malcolm Campbell	malcolm.campbell@plants.ox.ac.uk	Oxford University

Gene Expression & Microarrays

Peter Nilsson	nipe@biotech.kth.se	The KTH Genome Centre
Wout Boerjan	woboe@gengenp.rug.ac.be	Ghent University
Jörg Bohlmann	bohlmann@interchange.ubc.ca	University of British Columbia
Amy Brunner	Amy.Brunner@orst.edu	Oregon State University
Steve DiFazio	difaziosd@ornl.gov	Oak Ridge National Laboratory
John MacKay	john.mackay@rsvs.ulaval.ca	Université Laval

Informatics, Annotation & Database Development

Francis Martin	fmartin@nancy.inra.fr	INRA – Nancy
Jan Karlsson	jan.karlsson@plantphys.umu.se	Umeå Plant Science Centre
Dan Weems	dcw@ncgr.org	NCGR
Loren Hauser	hauserlj@ornl.gov	Oak Ridge National Laboratory
Nathalie Pavy	nathalie.pavy@rsvs.ulaval.ca	Université Laval
Pierre Rouzé	Pierre.Rouze@gengenp.rug.ac.be	Ghent University

Technical Approach

Selection of Species/Genotype

Of the many *Populus* species, and their numerous hybrids, that occur in ecologically important areas and/or are grown commercially, *P. trichocarpa* was selected because it is a major contributor to hybrid cottonwood germplasm in North America and worldwide, it is generally easily vegetatively propagated, and it occurs across a broad geographic range encompassing many ecologically important habitats.

Among all *P. trichocarpa* clones, 'Nisqually-1' was selected as the genotype that would be used in the shotgun sequencing effort. This individual is a native *P. trichocarpa* female clone first identified along the Nisqually River in western Washington and collected by Dr. Toby Bradshaw in 1995 and has been used to create several large F_1, F_2 and BC_1 generation pedigrees. It also had been used as the genomic template for one of only two BAC libraries for *Populus*. Nisqually-1 is freely available to the public via email requests to Dr. Toby Bradshaw, University of Washington, Seattle, WA (toby@u.washington.edu) and/or Dr. Max Cheng, University of Tennessee, Knoxville, TN (zcheng@utk.edu).

Sequencing Effort

The final aim of this project was to sequence the entire genome to an 8.5X depth based solely on shotgun approach. Out of this effort we determined that the genome size of *P. trichocarpa* is 485±10 Mbp.

Specifically, high-molecular weight nuclear genomic DNA was isolated from both apical meristems and root tips from vegetative propagules of Nisqually-1. Standard CTAB isolation protocols were followed for the apical tissue. The resulting template was contaminated with extra nuclear DNA (*i.e.*, chloroplast and mitochondrial DNA) and was therefore abandoned in favour of root tip tissues. A nuclei isolation protocol (*i.e.,* Percoll and cesium chloride density gradient centrifugation) was used to separate nuclei from all other cellular and subcellular components. DNA templates from both of these procedures were used to construct the following libraries: three ~3-kb plasmid libraries (two from the apex template and one from the root template), five ~8-kb plasmid libraries (all from the apex template), and two ~40-kb fosmid library (both from the root template). In addition, end reads were obtained from the 10X BAC library. Each library contributed to the sequencing depth based on the following:

- The actual insert size for the 3-kb plasmid libraries ranged between 2.5 and 3.0 kb. The sequencing effort from these libraries represented a 0.12X clone depth (or 59 Mbp insert coverage). Approximately 1,042,130 clones were used to obtain end sequence reads. Achieving an average of 650 bp reads at PHRED 20 or greater produced roughly 1.4 Gbp (ca. 2.8%) sequence coverage from the 3-kb libraries.

- The actual insert size of the 8-kb plasmid libraries ranged between 6.5 to 8 kb. The sequencing effort from these libraries represented a 28X clone depth (or 13.9 Gbp insert coverage). Approximately 1,910,000 clones were used to obtain end sequence reads.

Achieving 650 bp reads at PHRED 20 or greater produced roughly 2.5 Gbp (or ca. 5.1%) sequence coverage from the 8-kb libraries.

- The actual insert size of the fosmid libraries was 36 kb. The sequencing effort from these libraries represented an 18X clone depth (or 8.7 Gbp fosmid insert coverage). Approximately 243,000 fosmids were used to obtain end sequence reads. Achieving an average of 700 bp reads at PHRED 20 or greater produced roughly 340 Mbp (or ca. 0.71%) sequence coverage from the fosmid libraries.

- The actual insert size of the BAC libraries was 105 kb. The sequencing effort from this library represented an 8.6X clone depth (or 4.1 Gbp BAC insert coverage). Approximately 39,000 BAC were used to obtain end sequence reads. Achieving an average of 550 bp reads at PHRED 20 or greater produced roughly 43 Mbp (or ca. 0.09%) sequence coverage from the BAC library.

A total of 6.5 million sequence reads were obtained from all libraries.

Construction of a physical map

From the 10X Nisqually-1 BAC library created by Julia Vrebalov and Jim Giovannoni at Texas A&M University, 45,000 clones were fingerprinted with *Hind*III restriction enzyme and resolved on agarose gels. Restriction fragments were identified automatically using BandLeader software. Fingerprints were assembled into contigs with FPC software. Contigs were edited either manually or with automated tools. A total of 3,471 BAC contigs were created; 3,335 contigs were anchored to sequence assembly.

Construction of a high-density SSR genetic map for *Populus*

The genetic map of *P. trichocarpa* was estimated to be 2400 cM. A complete consensus genetic map in *Populus*, containing 637 SSR loci and 533 AFLP loci uniformly distributed across the genome at an average interval of 3.8 cM, was constructed. Based on end sequences of the BAC contigs, 3705 SSR primers were designed (Available: http://www.ornl.gov/sci/ipgc/), 672 of which were synthesized and used for segregation analysis. A subset of 53 progeny was genotyped in an effort to map these test-crossed loci in the maternal tree. Approximately 305 Mb of genomic sequence, containing in 162 scaffolds were assigned and oriented to the genetic map with 2 or more markers. Roughly 22,000 sequence scaffolds (ca. 180 Mb) remain unmapped.

EST, Full-length cDNA Sequencing and Gene Annotation

In order to train *ab initio* gene prediction algorithms existing EST libraries and newly created full-length cDNA libraries were consolidated or created, respectively. Approximately 250,000 *Populus* ESTs were assembled from existing libraries (e.g., xylem, phloem, cork, root, stress treatments, sapwood-heartwood transition zone, subtracted libraries, etc.). This EST resource was clustered into a 20,000 unigene set which contained over 4,000 *in silico* full-length sequences. The *in silico* set was manually curated and reduced to an 800 FL EST that was used

to train the *ab initio* gene prediction algorithms. The remaining ESTs were used to verify the predicted gene models.

Almost 1500 poplar full-length cDNA clones were sequenced in a collaborative project between the Genome Canada/Genome British Columbia Treenomix project in Vancouver, Canada, and the DOE Oak Ridge National Laboratory. Full-length enriched cDNA libraries were constructed in Vancouver from *Populus trichocarpa* Nisqually-1 using RNA isolated from developing xylem, green shoots and developing leaves, and bark based on the RIKEN protocol. Clones were selected through colony PCR based on the presence of an insert of greater than 350 bp. Library clones had an average insert size of 1.5 kb (range: 400-4500 bp). Clone ends were sequenced using the M13 forward and reverse priming sites of the pBluescript II SK+ phagmid vector used in library construction. Internal sequences were obtained via primer walking methods. Full-length cDNAs were assembled using PHRED/PHRAP and CONSED. Approximately 10,000 clones of varying lengths were analysed as single-pass reads, and roughly 1500 full-length sequences were used to train the gene algorithms. Sequencing was carried out at the Vancouver Michael Smith Genome Sciences Centre and at Oak Ridge National Laboratory.

Three separate gene prediction algorithms were used to create the basal annotation for the assembled genome. These were GRAIL-EXP at ORNL, EUGENE at the University of Ghent, and FgenesH at JGI. Each algorithm was trained based on the in silico and FL sets of cDNAs. In addition, Genewise was used to identify general open reading frames. All *Populus* EST data, along with EST data from other organisms (mainly Arabidopsis and rice), were used to validate the *ab initio* gene models. In total, each algorithm produced between 27,000-45,000 models. The filtered model set contains 58,000 gene models, with Eugene producing 31,045 unique models (39.8%), FgenesH producing unique 30,465 models (39.0%), GRAIL-EXP producing 11,179 unique models (14.3%) and Genewise producing 5,392 unique models (6.9%). Approximately 30% of all models have EST support. Gene models per linkage group ranges from 404 to 3149; 11,880 gene models remain in non-assembled, non-mapped regions of the genome.

Introduction

Forest trees are the dominant life form in many ecosystems. They provide structural and functional habitat for two-thirds of the Earth's terrestrial species and contain greater than 90% of all terrestrial biomass. Forests cover about 3.8 billion ha, or 30% of the global land surface. Managed and unmanaged forests provide external economic values through environmental benefits such as carbon sequestration, watershed protection, improved air quality and recreational habitats. In many parts of the world, forests furnish basic energy needs through the use of fuelwood (*e.g.*, in the Indian subcontinent fuelwood provides 60% of the energy). Among biologically derived materials, wood-based products are second only to maize in their monetary contribution to the U.S. economy. Moreover, approximately 25% by volume of all industrial materials are derived from forest-based resources.

Within the U.S. alone, the U.S. Department of Energy (DOE) estimates that with improvements in plant productivity and conversion efficiencies, 25% of the imported oil could be displaced by plantation-grown trees by 2050. Moreover, the conventional forest products industry annually contributes $400 billion to the world's economy. In sum total, forests contribute to the livelihoods of hundreds of millions of people worldwide. In addition to their economic and ecological importance, forest trees provide an opportunity to study biological processes not present in other plant systems. In contrast to agricultural crop species, forest trees are largely undomesticated. Forest trees contain adaptive gene complexes associated with long-lived perennial growth in wild populations, including natural populations of the oldest known organisms (*Pinus longaeva* & *Populus tremuloides*). Forest tree species allow study of population dynamics and adaptation on geographic and long-term temporal scales not available for annual or domesticated species. In addition, forest trees present unique forms of development, including the production of extensive secondary growth, resulting in the largest known organisms on the planet (*Sequoiadendron giganteum*).

Populus as a Model Perennial Woody Species:

The genus *Populus* has been adopted as a model for forest tree genetics. *Populus*, one of only two genera in the family Salicaceae, first occurred in the fossil record ca. 60 MYBP. Today, the genus consists of 5 sections, 30 to 40 species, and is circumpolar in the Northern Hemisphere. The majority of the species are obligately dioecious. Interspecific hybridization is readily achievable within single sections and among members of certain alternate sections of the genus. Hybrid vigor is common in interspecific crosses and has been exploited for establishing extensive commercial poplar plantations.

The genus *Populus* is especially well suited to serve as a reference genome for trees. *Populus* has 1) a small genome size -- the haploid genome size is ca. 485$\pm$10 Mbp, similar to rice, only 4X larger than *Arabidopsis* and 40X smaller than pine, 2) rapid juvenile growth - allowing meaningful measures of key traits to be taken within a few years after establishment of genetic trials, 3) ease of clonal propagation -- allowing destructive sampling without loss of the genotype, replication of experiments across time and space, and genetic stocks to be archived in clonal nurseries and 4) high-throughput transformation and *in vitro* propagation -- transgenic

trees can be produced, thus allowing detection and characterization of gene function. *Populus* is unique among tree genera in these combined traits.

***Populus* Genomics Resources:**

In May 2002, the DOE-JGI and ORNL in collaboration with Genome Canada-Genome BC, the Umeå Plant Sciences Center, and the University of Ghent, initiated the *Populus* Genome sequencing, assembly, and basal annotation project on the recommendation of the international *Populus* community (http://www.ornl.gov/sci/ipgc/sequencing_the_genome.htm). The 8.5X depth whole-genome shotgun data from *Populus trichocarpa* is now publicly available, completing the shotgun sequencing phase (http://genome.jgi-psf.org/poplar0/poplar0.home.html). The computational reconstruction of *Populus* genomic sequence suggests a whole genome size of ca. 485±10 Mb. Over 95% of the estimated 429 Mb of the euchromatic DNA is in the current assembly. The assembled genome has been reconstructed into large contigs and scaffolds, with typical size (N50) of 127 kb and 1.9 Mb, respectively. There are 58 scaffolds in the N50 count and 690 contigs. The genetic map and sequence assembly have been united using 691 mapped SSRs representing the 19 *Populus* chromosomes. These mapped SSRs occur on ca. 199 scaffolds and represent ca. 305 Mb of sequence. Single nucleotide polymorphisms occur 3 in every 100 bases. Ninety-one percent of the publicly available *Populus* EST sequences (ca. 200,000 individual sequences, forming ca. 11,885 unigene clusters and 12,759 singletons) were found in the assembled sequence and align at 85% identity over 50% of their length. A set of 2000 full-length cDNAs and 1112 full-length *in silico* cDNAs are being used to train three autonomous *ab initio* gene-calling algorithms: GRAIL-EXP, EUGENE and FgenesH. These algorithms have identified between 32,000 and 50,000 gene models in *Populus*. Of the 1112 *in silico* cDNA sequences, each over 1.0 kb, 95.0% showed a very high similarity (BLASTX score=200) to *Arabidopsis*.

Comparative Genomics: A conservative estimate gene order suggests that there are substantial regions of microcollinearity between the *Populus* and *Arabidopsis* genomes. Approximately 27% of DNA sequences from *Populus* BACs were homologous to protein-coding regions in *Arabidopsis*, and 46-58% of *Populus* gene pairs have pairwise homologs on a single *Arabidopsis* chromosome. In a BLAST comparison of the current *Populus* assembly with the *Medicago* BAC sequence database, we discovered that all 649 *Medicago* BACs had similarity to at least one *Populus* scaffold at E^{-10}. In a preliminary analysis, involving seven of the *Populus* scaffolds, the *Medicago* BACs mapped to the same *Medicago* linkage group more frequently than expected by chance (Poisson *P*-value < 0.05). In a BLAST comparison of the current *Populus* assembly with the *Medicago* BAC sequence database, we discovered that all 649 *Medicago* BACs had similarity to at least one *Populus* scaffold at E-10. On average 3.5 *Medicago* BACs contained high similarity to a typical poplar scaffold, with a maximum of 22 BAC BLAST hits for a single scaffold. The average identity was 88%, and the average length of hit was 108 bp. These preliminary indications of extensive homology between *Populus* and *Arabidopsis* and *Populus* and *Medicago* indicate that chromosome-level finishing, gap filling, and functional annotation of the *Populus* genome will benefit the *Medicago* and *Arabidopsis* communities, and *vice versa*.

Anticipated Impacts: The completed sequence from the *Populus* genome is an invaluable resource, not only to the forest research community, but also to plant biologists the world over.

However, just as the *Arabidopsis* and Human Genome communities have found, whole-genome sequence alone is of little utility to most researchers unless it is supplemented with high-quality annotation. To capture the wealth of information connecting plant genome structure, gene family diversity, and plant development, detailed knowledge of 1) genome structure and evolution, 2) gene architecture and 3) structural elements controlling transcription derived from known gene sequences, is essential. Thus, resolution of the genome structure through genetic mapping will be crucial to the accurate annotation of the sequence database. Designation of orthologs vs paralogs will be greatly enhanced by working hypotheses of genome evolution. Finally, functional genomics efforts, including the design of whole-genome microarrays and large-scale knock-out/knock-in mutagenesis programs, will depend on the accurate and complete annotation of the genome database. The multi-institutional science plan outlined below will: 1) facilitate future functional genomics studies by identifying and categorizing complete gene families in *Populus*, 2) foster comparative and evolutionary genomics work between three moderately related species with strongly contrasting characteristics: *Arabidopsis thaliana, Medicago truncatula* and *Populus trichocarpa* and 3) advance *Populus* biology related to the forest products industry, bio-based energy, phytoremediation and ecosystem sciences.

Genetic Resources

Panel Members: **John Davis,** Jörg Bohlmann, Stefan Jansson, Les Pearson, Barb Thomas, Chung-Jui Tsai and MarcVillar.

BACKGROUND AND SCOPE

A challenge faced by any scientific research community is to efficiently archive and distribute shared resources. Worldwide collaborative networks and reagent sharing have been a hallmark of *Populus* researchers for many years. The IPGC encourages continued active distribution of shared genetic resources in order to maximize the rate at which new discoveries in *Populus* biology are made.

Genetic resources include information such as genome sequence, EST sequences, genetic maps and physical maps. These types of resources are often best distributed electronically through web portals, and will be covered in greater detail in the Bioinformatics section of the Science Plan. DNA resources such as EST collections, large-insert libraries gene expression arrays, and DNA from mapping populations, and germplasm resources such as naturally occurring or conventionally bred genotypes, or transgenic plant material generated through functional genomics experiments may require distribution through stock centers.

CURRENT STATE OF THE FIELD

Physical Resources

Many groups and organizations around the world are making substantial contributions to research using *Populus* genetic resources. The archival and distribution of DNA resources is not unique for *Populus*, but no less challenging than for other species. In contrast, germplasm archival and distribution is potentially unique, and bears on achieving the IPGC's goal of archival and worldwide distribution of *Populus* genetic resources.

Conventional maintenance of forest tree germplasm requires outdoor space and maintenance. A major difficulty is that maintenance of populations and mapping lines derived from years of research continue to disappear due to lack of long-term support. Maintenance and distribution is performed *ad hoc* by individual researchers, but coordination among these individuals and cost-sharing by the community at large would be highly desirable. An example of coordination is illustrated by a *Populus* network for *ex situ* (= "off-site," archived at a location other than where the tree arose naturally) and *in situ* conservation (EUFORGEN network http://www.ipgri.cgiar.org/networks/euforgen/Networks/Poplars/PN_home.asp), which includes a database with genetic entries and a core collection of genotypes. An alternative to conventional germplasm maintenance is cryogenic storage (Tsai C-J and Hubscher SL. 2004. Cryopreservation in *Populus* functional genomics. *New Phytologist* 164: 73-81), which is being evaluated as an alternative to decentralized, land- and labor-intensive approaches to germplasm maintenance. In Europe, AFOCEL France has expertise in cryoconservation of *Populus* and other forest tree species (http://www.afocel.fr).

Financial Resources

Several valuable projects to help coordinate worldwide genetic resource use for *Populus* have recently obtained support, and we highlight some here. The *Populus* genome portal will integrate genome sequence information, create access to a fosmid tiling path across the entire genome, and generally integrate sequence and mapping resources to support functional and comparative genomic research (G. Tuskan). The International Populus Genome Consortium (IPGC) at Oak Ridge National Laboratories functions as a portal for many diverse genetic resources of *Populus*, serving as the community interface to the genome sequencing project, and containing a comprehensive list of links to other resources around the world (http://www.ornl.gov/sci/ipgc/). POPYOMICS helps coordinate the efforts of many groups in Europe to integrate genetic resources including ESTs, mapping information and germplasm (http://www.soton.ac.uk/~popyomic/index.htm; G. Taylor). Projects supported by Genome Canada help integrate efforts on genome and EST sequencing, microarray analysis, and germplasm (http://www.treenomix.com), (http://www.arborea.ulaval.ca/en/). Activation-tagged lines of *Populus* have been created through the Arborea project (S. Regan and J. MacKay), which will help create a larger resource in conjunction with a USDA-supported initiative in the U.S. (V. Busov and S. Strauss).

SCIENTIFIC OBJECTIVES

The scientific objectives being put forth here will collectively lead the *Populus* community to develop and support

1. A shared capacity for maintenance and distribution of germplasm, including mutant collections, pedigrees, transgenic trees & *in situ* populations

2. A shared capacity for maintenance and distribution of DNA resources such as BAC libraries, EST collections, and DNA from reference genotypes

Bioinformatics

The IPGC places high priority on establishing and/or partnering to establish a relational database for *Populus* molecular genetics that can be supported on a long-term basis. All activities related to the archival and distribution of shared genetic resources must be underpinned by a high quality, robust relational database with a public web browser interface. The database creates a forum for information exchange among community members, and serves as a nucleation point for supporting other community resources including stock centers.

DNA stocks (cDNA and genomic) and other molecular reagents

The IPGC aims to establish two *Populus* Biological Resource Centers, operated as parallel collections that closely coordinate the collection, maintenance and distribution of DNA stock resources and related molecular reagents. The rationale for centralization is to ensure consistency in material deposition, handling, distribution, and quality. As a general rule, it is expected that one Center will provide services to North America and one Center will provide services to the

rest of the world. The primary function of these Centers is to acquire, preserve and distribute DNA stocks and molecular reagents for use by the research community. It is anticipated that these Centers may distribute stocks on a cost-recovery basis. It is also anticipated that at least one or perhaps both Centers would be housed at institutions where DNA stock resources of other species are collected, maintained and distributed. In other words, the IPGC recognizes the economy of scale that can be accomplished if the *Populus* community can establish partnerships with existing stock centers, or planned centers that are designed to accommodate multiple species.

Germplasm

The IPGC supports a partially centralized approach to archiving and distributing germplasm, due to practical difficulties in centralizing all *Populus* germplasm resources. In contrast to Arabidopsis and most annual crop plants, which are propagated and distributed as seed representing accessions, cultivars or wild relatives, *Populus* germplasm is usually propagated and distributed in the form of dormant stem cuttings. Stem cuttings are placed in soil by the recipient, after which adventitious roots and pre-formed shoots grow and establish a new plant body that is the same genotype as (*i.e.*, a clone of) the donor plant. Propagation of genotypes by stem cuttings requires a good deal more space and effort than seed (which can be stored dry and in small containers). Usually genotypes of interest are maintained in field plots. Selected genotypes are sometimes naturally occurring genotypes from wild stands, although progeny sets from controlled, usually interspecific, hybridizations are often quite valuable resources for the community. Increasingly, transgenic materials are being produced that are valuable resources. The transcontinental range of several *Populus* species, and the transglobal range of the genus, creates practical difficulties in centralizing germplasm archival and distribution. Not all species and hybrids will survive and grow in a single location. Consequently it seems appropriate to distribute germplasm across a series of field plots around the world. Of course, over the long term, field sites are notoriously lost due to harvest, land sale, etc., and it is difficult to predict with any degree of certainty which sites would be most "stable" for archiving and distributing germplasm. Ownership category options to consider might be university, government, industry or even private land.

The IPGC recommends that a common registry of germplasm be established that would provide the community with a roster of available *Populus* material that is available for research. This registry would give the community notice of trial or plantation termination, thereby allowing others to obtain cuttings to continue specific lines. Furthermore, the IPGC recognizes the benefits of rapid and easy exchange of germplasm for research purposes. To this end, the IPGC encourages the development of blanket materials transfer agreement that is acceptable to all parties involved in *Populus* research. This agreement could be used routinely to document the current sites at which specific genotypes are being used for research, and to accelerate the transfer of scientifically useful germplasm among university, government, and industry research teams

The IPGC encourages the development of cryogenic storage methods for archiving and distributing germplasm. Cryo methods hold promise for reducing the space and labor requirements associated with maintaining field plantings. Cryo methods as applied to *Populus*

would involve freezing shoot tips after treatment with osmoprotectant compounds. Subsequent revival of the shoot requires some tissue culture steps, however when the methods are well defined for a plant species, they are relatively simple to perform. A centralized location for developing cryogenic storage and regeneration protocols for *Populus*, material deposition, cryostorage and distribution would be desirable, with potentially several other "inactive" storage facilities elsewhere.

SUMMARY

The *Populus* community is not unique in its need to establish some centralized entities to allow sharing of genetic resources. In particular, archival and distribution of DNA resources could best follow the centralized model adopted by many communities including Arabidopsis. To be successful these Centers should be underpinned by an excellent and robust relational database, and opportunities for partnerships with other communities of researchers should be pursued in order to reduce costs. The *Populus* community is somewhat unique in the challenges of sharing germplasm, since the germplasm is propagated vegetatively. A common registry and transfer agreement should help facilitate sharing, albeit in a decentralized way. Finally, the development of cryostorage methods is encouraged in order to reduce the overall costs, and increase the overall efficiency of germplasm archival and distribution.

High-throughput Phenotyping

Panel Members: **Mark Davis**, Stephen Kelly, Bob Kodrzycki, Laigeng Li, Gary Peter, Simon Potter, Lacey Samuels, Steve Strauss, Bjorn Sundberg, Gail Taylor and Stan Wullschleger.

BACKGROUND AND SCOPE

Considerable progress has been made by the plant science community in the development of techniques for generating populations of mutant or transgenic organisms. Activation tagging, gene/promoter trapping, fast-neutron deletion, and RNA silencing have all been shown effective in creating loss- and gain-of-function mutants in higher plants. Application of these techniques has led to the development of large populations of organisms with altered patterns of gene expression. These populations are proving to be extremely useful in a wide variety of studies aimed at better understanding the role of specific genes, gene families, and gene products (i.e., proteins) in plant biochemistry, physiology, and growth and development.

SCIENTIFIC OBJECTIVES

Considering the size of current transgenic populations, and recognizing that capabilities for generating even larger numbers of transgenic lines will undoubtedly be enhanced in a post-genomics world, scientists are finding it necessary to also develop rapid and reproducible methods to identify changes in plant characteristics (i.e., phenotype). Such information will be critical if we are to associate altered phenotypes with the presence or absence of a particular gene, and therein to determine gene function for the thousands of genes present within the poplar genome.

With this as one of the principle goals of the International *Populus* Science Plan, here we identify the tools and technologies required for the high-throughput characterization of altered plant phenotypes in trees. Our aim is to identify potential approaches that while still in their infancy might ultimately serve as high-throughput tools for assessing gene function in trees. Our emphasis lies in the development of new and rapid approaches for quantifying variation in wood and/or tree-specific anatomical, morphological or physiological traits as part of a future functional genomics effort.

Scientific Objective #1

Develop phenotyping tools for traits of interest to the tree biologist.

Wood quality

What is the range of variation in wood quality phenotypes for selected genotypes, clonal material and control transformed *Populus* trees? What are the appropriate ages for wood quality determinations? What genes affect wood phenotypes? What are the molecular mechanisms that underlay the variation in wood quality phenotypes? What controls the developmental mechanisms that affect wood quality phenotypes?

Short-term Goals:

- With existing methods, (Molecular Beam MS, Fiber Quality Analyzer, Silviscan II, Microcat Scan) determine a baseline and variation in wood quality phenotypes (tip to base: pith to bark) for pedigrees in breeding populations, clones and Nisqually 1 tissue culture regenerated lines for multiple ages;
- Determine appropriate age(s) of trees for reliable comparisons of wood quality phenotypes useful for detection of altered phenotypes.
- Develop reliable tests for multiple wood quality property determinations from young single seedlings;
- Develop in forest methods to accurately determine wood quality phenotypes for standing trees, e.g., via Near Infrared Spectroscopy.

Mid-term Goals:

- Develop associations between wood quality phenotypes and changes in gene expression;
- Screen populations of transgenic lines for altered wood phenotypes: with multiple methods: in forest, nondestructive and destructive sampling;
- Understand the molecular basis of tension wood formation;
- Identify the functions of all carbohydrate biosynthetic genes;
- Identify genes that affect fiber and vessel element morphology: fiber length, microfibril angle, cell wall thickness, wood density, # fibers/unit volume.

Long-term Goals:

- Isolate and obtain proof of genes that underlay wood quality QTLs;
- Understand the role of allelic interactions in controlling wood quality traits;
- Understand interactions between growth rate and juvenility changes with wood quality phenotypes.

Wood anatomy

Science Questions:

Wood anatomic phenotypes would be one of most exclusive characteristics distinguishing woody plants from others. It is also the wood anatomic features that make Populus genome program so unique, irreplaceable, and valuable to plant biologists as well as forest scientists. To understand functional genomics of how plants form wood tissues, it is essential to develop practical phenotyping protocols and to phenotype wood anatomy in various mutants, genotypes, and major Populus species. Currently, the wood anatomy of Populus species is rather understudied. In the era of genomics, one ultimate question that would be answered is how wood anatomic phenotypes can be bridged up with the knowledge of functional genomics that the Populus Genome Science Plan is targeting .

Short-term Goals:

- Establishment of the standard protocol(s) for wood anatomy phenotyping, in juvenile and mature stage; workshops are needed;
- Systematical documentation of wood anatomic phenotypes in major Populus species under the standard protocols.

Mid-term Goals:

- Phenotyping of the collections of Populus mutants and genotypes;
- Identification of genes corresponding anatomic phenotypes through forward or/and reverse genetics.

Long-term Goals:

- Characterization of all genes that are responsible for various wood anatomic phenotypes in Populus genome;
- Comprehensive understanding of functional genomics of wood anatomic characteristics.

Scientific Objective #2

Associate traits of interest as detected in phenotypic studies to gene expression, metabolic characterization, and protein profiling.

Molecular Phenotyping: Need to do microarray analyses of gene expression on well characterized pedigrees and mutants identified in early- and field-detection screens.

Scientific Objective #3

Greenhouse, laboratory, and field-level screening of aberrant phenotypes will generate vast amounts of data that, in order to a valuable community resource, must not only be archived, but also must be accessible.

Database: Databases of phenotypic information from baseline studies and transgenic populations need to be built and accessible to the community. Linkages to molecular phenotyping, metabolic profiling, and protein characterization also needs to be organized

SUMMARY

Our need for high-throughput phenotyping is certainly not new. Long before the genomic era, techniques for assessing plant phenotypes were needed as new cultivars and/or varieties were being developed through conventional breeding approaches. Today, however, tools specific to the high-throughput screening of plant phenotypes offers to promote understanding of gene function and will soon become a valuable component of the plant genomic toolbox. Although the scientific community currently relies heavily on visual evaluation of mutants, imaging techniques that allow immediate and non-invasive detection of altered plant characteristics, before visual symptoms appear must be developed for monitoring changes in phenotypic characteristics of transgenic plants expressing those genes.

Physical Mapping & Marker Development

Panel Members: **Patricia Faivre-Rampant**, Rishi Bhalerao, John Carlson, Carl Douglas, Berthold Heinze, Maurizio Sabatti and Jerry Tuskan.

BACKGROUND AND SCOPE

Genetic mapping and marker development and use in a traditional sense have been employed for decades in forest tree research. This has been true for many *Populus* species as well. Markers in the form of allozymes, randomly amplified polymorphic DNA (RAPD), restriction length polymorphisms (RFLP), amplified fragment length polymorphisms (AFLP), simple sequence repeats (SSR), and sequence tag sites (STS) have been used in genetic fingerprinting exercises, linkage studies, quantitative trait loci (QTL) experiments, studies of marker variation and introgression in natural populations, and consensus mapping across pedigrees. Following the completion of genomic sequencing, mapping and markers will play a role in assembling the draft sequence into a final assortment of 19 chromosomes, in creating a consensus map across *Populus* species, in marker-assisted selection, and in conducting association studies between anonymous markers and adaptive genotypes.

A community-wide, post-genomic mapping and marker development effort will allow progress towards an understanding of the *Populus* genome structure. It will provide the development of genomic tools needed to identify genes and their position along the chromosomes. Availability of reference maps and marker sets will contribute directly to elucidate gene function via comparative mapping and QTL detection. Development of new markers, in the form of single nucleotide polymorphisms (SNP) for example, will provide a suitable forum for the assessment of genetic variation in natural *Populus* populations. A key strength of *Populus* as a model organism for genomic research is its occurrence across broad environmental gradients in locally adapted native populations. In additional, comparative mapping between several species may shed light on species differentiation and genome organization variation. A set of reference markers could also be used to clarify the evolutionary history of the genus *Populus* and its closest relatives.

CURRENT STATE OF THE FIELD

Physical Resources

There are numerous and various laboratories worldwide that have current capabilities in marker development, mapping algorithms, and fingerprinting. These capabilities include high-throughput sequencing and fingerprinting capacity, colony picker for manipulation of BAC libraries, robotic pipeting and solution handling, software and hardware required for referencing and analyzing the resulting data, and extensive pedigrees. These resources create a high-throughput platform for genotyping, mapping, detection and isolation of genes of interest.

Financial Resources

Most of the progress in marker development and mapping in *Populus* has come through individual investigators receiving stand-alone funding for single foci experiments. The trend in funding over the past five years however has been towards larger, more integrated national or multi-national projects involving several laboratories. As successful as these efforts have been, the worldwide *Populus* community is comparatively under funded relative to other plant genomics communities. Creating a shared mapping and marker resource for the entire *Populus* community will require greater integration and investment in these activities than has been historically available.

Financial resources and support of post-genomics mapping and marker approaches currently limits their application and use in *Populus*. Limited financial resources impacts future research efforts by restricting the number of scientist, technicians, and students working in the field and by restricting the scope of on-going experiments. These two resources represent the principal limitation to post-genomic development of consensus maps among species, transfer of genomics information among species, association studies for the discovery of adaptive traits, and ultimately the identification, elucidation and isolation of novel genes and gene regulation via de novo mapping and marker development.

SCIENTIFIC OBJECTIVES

The scientific objectives being put forth here will collectively lead the *Populus* community to develop a shared capacity and resource development related with the following areas:

1. Genetic mapping
2. Physical mapping
3. Development of new markers

Genetic Mapping

Over a half a dozen genetic maps, based on an assortment of marker types, are already available for several pedigrees and species, including *P. trichocarpa*, *P. deltoides*, *P. nigra*, and *P. alba*. These genetic maps have been generated for diverse purposes (e.g., genome organization, QTL detection). To date, only one genetic map constructed for a hybrid *P. trichocarpa* x *P. deltoides* represents a complete genetic map. Further, a common chromosome nomenclature is proposed nevertheless a select set of common molecular markers have not been created. A reference map and marker set would allow the integration of all current and future genetic maps and would accelerate the process of gene identification and isolation.

Short-term goals:

- Develop a reference set of microsatellite and ESTP markers for comparative genetic mapping between all members of the *Populus* genus.
- Establish a framework genetic map using a reference segregating pedigree and microsatellite markers that saturates the complete genome.
- Establish a database in connection with the Informatics, Annotation & Database Development panel.

Mid-term goals:

- Create a consensus map in all relevant species.
- Develop fine-scale genetic maps based on microsatellite and AFLP markers for the localization of genes uniquely related to woody, perennial plant growth and development (e.g., disease resistance, tolerance to abiotic stress, wood formation).
- Develop genetic maps including genes in conjunction with EST and physical mapping research to identify and isolate regulatory elements responsible for important traits.

Physical Mapping

Bacterial artificial chromosome (BAC) libraries could provide a reservoir of genetic information on the *Populus* genome. Several BAC libraries have been produced in *Populus*, including one for *P. trichocarpa*, one for *P. deltoides* x *P. nigra* and *P. tremuloides*. Such libraries can be used as a sequencing substrate and/or to develop a physical map for the *Populus* genome that could be used for locating genes of known function. BAC-end sequence can be used to anchor draft sequence to physical genome structure. New markers from BAC sequence would be used to unite the genetic map and physical map in *Populus*. BAC sequence and expressed sequence tags (EST) analyses could also provide data for gene-specific primer designation.

Short- term goals:

- Sequence BAC ends for libraries for the major species from 5 of 6 the *Populus* sections (Aigeros, Leucoides, Populus, Tacamahaca, and Turanga).
- Initiate physical maps based on a combination of BAC fingerprinting, BAC ends and EST analysis.
- Establish a database in connection with the Informatics, Annotation & Database Development panel.

Mid-term goals:

- Integrate the physical map and the genetic map obtained for the consensus map using SSRs and EST-P.
- Develop cytogenetic maps that assign BAC contigs onto *Populus* chromosomes by FISH or other physical mapping approaches.
- Develop a model for the organization and evolution of *Populus* chromosomes that incorporates information from each of the above approaches.

Long-term goals:

- Use the *Populus* genomic sequence to conduct homologous, heterologous and anonymous queries for comparative studies with Arabidopsis and rice databases.
- Identify sequence homology of genomic regions with Arabidopsis and rice for the identification of polymorphic and *Populus*-specific regulatory sequences.

Development of New Markers

Direct selection for desired genes will be more efficient than selection based on phenotypes when environmental effects are important. Furthermore, evaluation of genetic diversity will be more effective with molecular markers corresponding to the gene responsible for adaptive traits. Thus, identifying polymorphism within genes controlling growth, adaptation and development is very important. To achieve this objective a validation of the gene influence on the trait can be tested by linkage disequilibrium studies. Alternatively, EST sequence and analyses can provide data for gene-specific primer designation and the development of SNPs. These approaches will provide databases that facilitate high resolution of QTL, marker-assisted selection and genetic improvement. Additionally, anonymous SNP markers can be used to characterize differences in natural populations along environmental gradients in an effort to identify markers associated within adaptive traits.

Short-term goals:

- Identify candidate genes for tree-specific growth, development and adaptive traits in collaboration with research efforts in genetic physical mapping and functional genomics.
- Design of new primers that amplify genes responsible for tree-specific growth, development and adaptive traits.
- Sequence homologous regions within the *Populus* genus to identify variant genes and regulatory sequences.

Mid-term goals:

- Develop and design a universal gene-primer set using synteny between *Populus*, Arabidopsis and rice databases for comparative studies.
- Designate a common marker set for germplasm assessment, cultivar identification, genetic relationship studies and diversity studies.

SUMMARY

If implemented, this five-year science plan will lead to an integration of the physical maps and genetic maps in *Populus*. This plan will also aid in the characterization and isolation of genes responsible for traits of adaptive and agronomic importance in *Populus*. The development of mapping tools will aid breeders in creating new clones combining various selected traits. Therefore, the benefit of the application of genomics to poplar will contribute to increase the precision of forest management. This plan will facilitate the description of genome organization of *Populus* relative to other model species. And finally, the knowledge generated from studying the *Populus* genome will therefore be applicable to related species (willow) and other woody species. In this way, the potential for micro evolutionary studies between closely related species, and between populations within species, will be created.

Tissue Culture & Transformation

Panel members: **Rick Meilan**, Malcolm Campbell, Shujun Chang, Takashi Hibino, Ove Nilsson and Gilles Pilate.

BACKGROUND AND SCOPE

Although scientists often generically refer to the procedure by which plants are genetically engineered as "transformation", it really entails two separate processes. The first, transformation, involves stable integration of DNA into the chromosome of an individual plant cell. The second step, regeneration, is the process by which individual transformed cells are coaxed, through hormonal and cultural manipulations, to develop into a whole plant. Generally, regeneration is the most limiting of the two processes.

Transformation is essential for introducing novel traits, particularly those that are not available in the gene pool of sexually compatible species. In addition, transformation is an important tool for the analysis of gene functionality. By constitutive or conditional up- or down-regulation (knock-in/knock-out, KI/KO) of the target gene, important information about its function and downstream targets can be obtained. Constitutive expression from the cauliflower mosaic virus (CaMV) 35S promoter is still the dominant approach, but there is a growing demand for more refined methods using tissue- and temporal-specific promoters. Few other promoters have been characterized and used in the poplar system; however, promoters active during various stages of wood formation or floral development are in high demand, both for basic and applied research. Inducible systems are also very important tools for the study of gene function. We lack a good understanding of how the systems that have been developed in Arabidopsis perform in poplars. It is especially important to have reliable inducible systems for use with transcript profiling or proteomics, in order to identify early downstream targets of transcription factors.

Finally, a versatile system for long-term storage of poplar genotypes is critically needed. Plant preservation has traditionally relied on seed banks or collections in gardens, orchards, or clone banks. Seed storage is the most common mode of preservation for many angiosperms and gymnosperms, but for species with long generation times, actively growing samples are the preferred form of preservation. However, maintenance of actively growing collections is labor-, time- and space-consuming, not to mention the potential for pest infestations and microbial contamination. In addition, plants or tissues maintained ex vitro or in vitro are subject to growth constraints and environmental fluctuations that can reduce viability and inhibit propagation potential. Field planting transgenic trees for long-term germplasm preservation suffers from the additional burden of regulatory constraints, ecological and security concerns, and difficulties in maintaining juvenility. Cryopreservation is viewed by many as the only viable method for long-term storage and distribution of important germplasm stocks.

CURRENT STATE OF THE FIELD

While reliable transformation systems have been developed for pure species and hybrids in the section *Populus*, the genotypes in other sections have been found to be recalcitrant. To date, only a limited number of genotypes in sections Tacamahaca and Aigeros have been successfully

transformed. Most notably the genotype being sequenced, Nisqually-1, which is in the section of Tacamahaca, has been very resistant to regeneration. Regardless of section, the most commercially and experimentally important genotypes tend to be the least transformable.

Transformation barriers can be divided into two broad categories: biology and maturity. Biological barriers have not received much attention, but they are likely the reason that many clones with good organogenic potential are not transformable. Maturation-related obstacles, on the other hand, have been well documented as a major limitation in the transformation of many important genotypes. However, rootability is often confused with regenerability. In the case of P. deltoides, commercial clones are screened for their rootability before clonal testing. As long as cuttings root well, axillary buds flush and the propagation cycle is complete. However, in vitro regenerability is dependent on the ability to induce adventitious shoots from cells. Several studies have shown that seedlings from recalcitrant clones can be easily regenerated. The apparent lack of hereditary control implies that the decline in regeneration potential is related to maturity. In another case, leaf explants from a recalcitrant clone were easily regenerated after a year of in vitro culture on medium containing cytokinin. The regenerated plants turned out to be highly organogenic. This suggests that the recalcitrance was related to maturation. Comparing expression profiles in tissues that differ in regeneration potential, after growth on regeneration medium, could help identify genes involved in regulating the process. In addition, gene expression profiling done on tissues gathered during the juvenility-to-maturity transition could help identify genes affecting regeneration.

During transformation, a selectable marker gene is usually introduced simultaneously with the gene of interest. This marker gene is often one that imparts resistance to an antibiotic or herbicide. As a result, a transformed cell can be isolated on a medium containing the appropriate selection agent. While this method is convenient, it can be problematic. First, performing subsequent rounds of transformation may not be possible because there are a limited number of selectable marker genes available. Second, various selection agents can dramatically affect transformation efficiency. Finally, the presence of a selectable marker gene is an impediment to gaining public acceptance of genetically engineered plants.

Recently, alternative selection systems have been developed. One such system is based on the growth medium lacking a substance needed for proper metabolic activity or physiological development. A particularly attractive option is based on the inability of a cell to regenerate a whole plant without the addition of a phytohormone or phytohormone derivative to the culture medium at a precise step in the regeneration process. This is true for most plant transformation protocols, which rely on the addition of cytokinin to the medium for inducing the differentiation of adventitious buds from transgenic callus. This supplement has to be transitory because cytokinin can have an inhibitory effect on subsequent root development.

New vectors have been developed that carry a gene critical for the biosynthesis of cytokinin from adenine. Cells transformed with these vectors will continuously produce cytokinin; only they will be able to regenerate into shoots, whereas untransformed cells will remain in an undifferentiated state. Further, these new vectors carry an excision system for removal of the cytokinin biosynthetic gene once transgenic shoots have been regenerated, allowing for rooting of these shoots. Excision systems consist of a recombinase gene, often under the control of a

reliable inducible promoter, and recognition sites flanking the DNA fragment to be removed. These alternative selection systems are attractive because they: 1) have the potential to increase both the yield and speed with which transgenic poplars can be produced, facilitating the use of various high-throughput strategies, such as KI/KO, and 2) may eliminate the need for both specific selection and regeneration conditions, making it possible to transform a wider array of genotypes.

SCIENTIFIC OBJECTIVES

This portion of the Science Plan has two overall objectives: 1) improving *Populus* transformation/regeneration efficiency and 2) developing tools to increase the utility of *Populus* as a model plant system. Thus, under each of two Scientific Objective headings below, there are several numbered sub-tasks.

Scientific Objective #1

Develop a more robust transformation protocol, applicable to a wide array of genotypes, by improving our understanding of the fundamentals underlying transformation and regeneration. This system should allow for significant reduction in the time needed to regenerate transgenic poplars. The ultimate goal is to develop a high-throughput transformation protocol with over 50% transformation efficiency and rapid turnover (e.g., co-cultivation to soil in 3-4 months).

Highly Efficient *Populus* Transformation

It is essential to develop a highly efficient transformation protocol in order to over- or under-express single or multiple genes. The key is to be able to rapidly test a large number of constructs in the shortest time possible. Several aspen clones such as 717-1B4 (*P. tremula x P. alba*) and a few *P. tremula x P. tremuloides* clones, which have transformation efficiencies equivalent to that of tobacco, can serve the gene-testing purpose. However, their genetic distance from Nisqually-1 and their restricted planting ranges could possibly limit the utility of these clones for high-throughput analysis of gene functionality or field-testing.

Short-term goals:

- Agree on the genotypes to be used
- Select a clone that is closely related to Nisqually-1 and has a transformation efficiency similar to that of tobacco (another clone entirely or progeny of Nisqually-1)
- Select a relatively early-flowering genotype (until better clones or systems are identified) not only for testing flowering-related genes, but also as a method for testing other genes and producing viable transgenic seeds
- Improve the transformation efficiency of the genotype being sequenced, Nisqually-1, which could also be used as the recalcitrant clone for testing regeneration- and transformation-related genes
- Identify poplar genes that interfere with or promote regeneration
- Identify poplar genes related to transformation

Mid-term goals:

- Test (over- or under-express) various genes thought to affect regeneration and transformation
- Develop a better gene-testing model system, which has consistently high transformation efficiency and a shorter infection-to-soil cycle (similar to that of tobacco)
- Develop an efficient gene-testing model with an early-flowering genotype

Future applications:

- Produce a complete collection of lines (saturate the genome) through: activation tagging, enhancer or gene trapping, KI/KO, and insertional mutagenesis
- Develop the publicly accessible database for the collection of transgenic lines and link the information with functional analysis and sequence data
- Achieve a global understanding of the genes related to morphogenesis, maturation, rooting, and Agrobacterium susceptibility

Universal Vectors

It would be useful to have some universal shuttle and binary vector backbones for assembling both over-expression and KO constructs. However, these progenitor constructs need to be freely available, which is now frequently not the case (e.g., recent reports of difficulties with agreements governing the use of vectors that are commonly utilized in the assembly of RNAi vectors).

Short-term goals:

- Evaluate usefulness of resources developed by the Arabidopsis community
- Mid-term goals:
- Assemble shuttle and binary vector backbones for over-expression and suppression
- Future applications:
- The ability to easily produce constructs (high throughput) needed to verify the function of individual genes through the use of KI/KO constructs

Transformation without Tissue Culture

The goal is to produce an in planta transformation method similar to the one that has been established for Arabidopsis.

Short-term goals:

- Establish a basic transformation protocol of using reporter genes

Mid-term goals:

- Optimize the protocol by testing various transformation parameters

Future applications:

- A faster, more efficient transformation system that has lower risk of somaclonal variation than traditional methods

Scientific Objective #2

Develop molecular tools that will allow *Populus* to become an even better model species and a more attractive system for doing comparative, developmental biology research.

Inducible Promoters

Plant genetic engineering has largely been performed using strong constitutive promoters, the aim being to obtain maximum levels of expression and, consequently, maximal effect. Such a strategy may be useful for imparting certain commercial traits (e.g., insect resistance or herbicide tolerance), but is not practical when the aim is to alter the expression of an endogene. Indeed, numerous cases of gene silencing have been reported. In addition, altering the expression of key genes, such as transcription factors or other regulatory genes, may have lethal or at least strong negative effects on plant development. For the IPGC, the major application of genetic engineering is large-scale evaluation of gene function via KI/KO strategies. The production of transgenic plants altered in the expression of numerous genes vital to plant growth and development will be impossible unless a reliable system exists for conditional transgene expression. It is imperative for the inducer to be highly specific to the target promoter and lack phytotoxicity. It is also important for the target promoter to be tightly regulated (i.e., no "leaky" expression) and that high-level expression is conferred upon induction.

Short-term goals:

- Obtain existing constructs for various systems (ethanol, dexamethasone, estrogen, etc.)
- Transform constructs into *Populus*

Mid-term goals:

- Test for toxicity, inducibility, specificity of inductive agent
- Determine the level and specificity of expression

Future applications:

- The ability to over- or under-expression selected genes that may be lethal or have otherwise undesirable consequences

Gene Excision

In any genomics project, the availability of a reliable excision system is valuable both for applied and fundamental studies. An excision system will not only permit removal of selectable marker genes, to alleviate public concern, but it will also allow for easy retransformation using vectors derived from a common backbone. Moreover, the alternative transformation systems described above depend on a reliable excision system.
Because transposon systems have proven difficult to control, alternative excision systems were developed. These systems exploit naturally occurring recombinases, such as cre/loxP or R/RS. Gene excision vectors include: 1) a recombinase gene under the control of an inducible promoter, and 2) recognition sites that flank the targeted DNA. The goal of this project is to

evaluate various excision systems that are available in order to determine which is the most appropriate for use with poplar. For each system, it is important to assess the efficacy of induction, the level of inducibility, and leakiness of the promoter (as described above). In addition, it is necessary to ascertain the efficacy of the recombinase, and how clean excision is, in various poplar genotypes.

Short-term goals:

- Obtain existing constructs for various systems
- Transform constructs into *Populus*

Mid-term goals:

- Evaluate transgenic plants for clean excision of target sequence

Future applications:

- The ability to selectively remove unwanted genetic material (e.g., selectable markers)

RNA Interference

Because poplars have a long juvenile period and are dioecious, classical mutagenesis or insertional mutagenesis are not practical tools to study loss of gene function in poplar. The poplar community is, therefore, dependent on indirect methods, such as anti-sense RNA or RNA interference (RNAi), for down regulation of gene expression. Due to the dominant nature of this approach, and the relative ease with which large numbers of RNAi constructs can be generated, it should be possible to design large-scale programs with the ultimate goal of down-regulating all poplar genes. Several approaches to RNAi have been described, but these need to be thoroughly tested in poplar.

Short-term goals:

- Test various approaches to RNAi for efficiency
- Determine the extent to which similar gene family members can be targeted individually or collectively (specificity)
- Test the use of tissue- and temporal-specific promoters to drive RNAi constructs
- Test the use of inducible RNAi constructs (especially important for the analysis of genes whose loss-of-function prove to be lethal or in other ways prevent in vitro regeneration, and in combination with transcript profiling and proteomics approaches)

Mid-term goals:

- Evaluate stability over time (i.e., will the silencing be maintained over several cycles of growth and dormancy?), which is especially relevant for perennial plants
- Development of stock centers where RNAi constructs for all poplar genes can be deposited and shared with the rest of the poplar research community

Future applications:

- The ability to effectively silence any gene in a given tissue or at a specific point in time
- Development of clonal archives of transgenic plants where all individual poplar genes have been targeted

Floral Induction

Most genotypes of poplar have a five- to seven-year juvenile period. This is a serious impediment for the many researchers working on flowering control. In order to make more rapid progress in this area, an early-flowering genotype or a simple, reliable method of floral induction is needed.

Short-term goals:

- Conduct an exhaustive search for genotypes that are known to flower early
- Experiment with various inductive treatments (chemical, physical, cultural) that have been shown to be successful with other woody angiosperms
- Identify genes involved in controlling the timing of floral development

Mid-term goals:

- Transform putative flowering-time genes into normal-flowering genotype(s)
- Verify that induced flowers are functional

Future applications:

- An important tool for breeding and marker-aided selection
- With a highly efficient early-flowering system, it would even be realistic to generate homozygous insertional mutants
- It might also be possible to generate populations for rapid mapping of a specific trait

Cell Culture

Synchronized cell culture systems, such as tobacco BY2 cells, in which it is possible to obtain an absolutely homogenous response to added compounds or treatments, is an important tool for studying the regulation of growth and development.

Short-term goals:

- Further development of stable *Populus* cell cultures that are highly homogenous
- Develop methods to efficiently induce differentiation of tracheary elements with a high degree of synchronization

Mid-term goals:

- Development of a highly synchronized system for the study of cell-cycle regulation

Future applications:

- The possibility of inducing relatively undifferentiated cell cultures to differentiate into specific cell types, as with the *Zinnia* tracheid development system, is a powerful tool for analyzing developmental transitions and cell-cycle regulation

Virus-Induced Gene Silencing

Virus-induced gene silencing (VIGS) has emerged as a powerful approach to modify gene expression in herbaceous annual plants. Several viral vectors have been described, which function to silence endogenous genes within a range of plant genomes. VIGS is a systemic and epigenetic phenomenon, which means that gene silencing occurs throughout the plant, but is not transmitted from generation to generation. Due to the dominant nature of this loss-of-function approach, and the relative ease with which large numbers of viral constructs can be generated, it should be possible to design large-scale programs with the ultimate goal of down-regulating all poplar genes via VIGS. Several approaches to VIGS have been described, but these need to be thoroughly tested in poplar.

Short-term goals:

- Test various VIGS vectors for efficacy with poplar
- Development of new VIGS vectors for poplar, based on tree viruses, such as poplar mosaic virus (PopMV)
- Determine the extent to which similar gene family members can be targeted individually or collectively (specificity)

Mid-term goals:

- Evaluate stability over time (i.e., will the silencing be maintained over several cycles of growth and dormancy?), which is especially relevant for perennial plants
- Development of stock centers where VIGS constructs for all poplar genes can be deposited and shared with the rest of the poplar research community

Future applications:

- The ability to effectively silence any gene in a given tissue or at a specific point in time
- Development of clonal archives of transgenic plants in which individual poplar genes have been targeted

Cryopreservation

It is vital to develop and test protocols to enable transgenic stocks to be stored at low cost and with low rates of somaclonal variation. Cryogenics is the most logical means for long-term storage of transgenic germplasm (especially KO/KI lines). Vitrification is a simplified cryostorage procedure that eliminates the need for expensive cooling devices. It relies on the controlled application of cryoprotective solutions that desiccate and penetrate cells at low temperatures, leading to increased viscosity of the cytosol and formation of an immobilized solution state. When vitrification is optimally achieved, even the most delicate tissues can

survive direct immersion and storage in liquid nitrogen without ice crystal damage. Because of its simplicity and low cost, vitrification has become a preferred method in recent years.

Short-term goals:

- Optimize the key steps: 1) conditioning, 2) pre-culture regimes, 3) cryoprotection, 4) warming, and 5) recovery regimes
- This will include testing factors known to affect successful storage and recovery such as sorbitol concentration in the pre-culture media, composition of vitrification solution, and PGRs in the recovery media, etc.

Mid-term goals:

- Evaluate genetic fidelity of re-established plants (e.g., by using morphology, NIR spectra, and simple sequence repeat (SSR) markers previously developed for *Populus* spp.)
- Develop inventory-tracking software (e.g., to record species or genotype origin, genetic modifications, number of propagules stored, cryostorage parameters, viability, recovery regimes, micropropagation procedures, genetic fidelity, and other characteristics)
- Establish a mechanism to fund a long-term storage and distribution system

Future applications:

- The infrastructure to maintain and distribute important stocks to the poplar research community.

SUMMARY

Implementation, this portion of the Science Plan will increase the efficiency with which the poplar research community can verify gene functionality via transgenesis. It will also result in the development of systems for storing, distributing and recovering existing and newly developed germplasm. Utilization of the tools developed through this work will lead to a better understanding of the way in which tree growth and development is regulated. Ultimately we will be able to readily domesticate trees to help satisfy society's ever-increasing demand for renewable forest resources.

Metabolic Characterization & Metabolomics

Panel Members: **Tim Tschaplinski**, Thomas Moritz, Andrea Polle, Scott Harding, Janice Cooke and Reinhard Jetter.

BACKGROUND AND SCOPE

The emerging science of metabolomics couples metabolite profiling with the analysis of mutant and transgenic lines to elucidate protein function, the structure of metabolic pathways, and offers tremendous potential to discover and assign function to novel genes. Stated explicitly, metabolic profiling is the unbiased, relative quantification of the broad array of cellular metabolites, and their fluxes. As such, metabolic profiling can provide information on how gene function affects the complex biochemical network, and the levels of regulation of biochemical networks that are not revealed by DNA microarray technology. A comprehensive functional genomics research platform, that links metabolite profiling to gene expression arrays and protein profiles, will facilitate the cataloguing of genes. About 500-1000 metabolites may be expected to accumulate to detectable levels in a typical eukaryotic genome, which codes for >10,000 proteins. Therefore, it is unlikely that a single gene knockout or up-regulation event will often lead to direct relationships of a single gene completely regulating the production and accumulation of a single metabolite. Some examples indicate that genetic mutations can lead to changes that are highly pleiotropic, depending on where the mutation is operating in the metabolic networks. However, the ability to detect a wide array of metabolites (and their fluxes) will permit determination of how biochemical networks, with their distributed control (regulation), have been perturbed.

Successful deployment of metabolic profiling requires the development of rapid, reliable, and efficient assays for detecting phenotypes that are metabolic variants within natural or mutated populations. Assays need to be developed which will allow the detection of as many metabolites as possible and preferably at high-throughput rates. Although the desire is to have a single analysis that captures all metabolites in a short time, there is, as yet, no single "silver bullet" analysis that will be appropriate for all metabolites with a high degree of sensitivity and resolution. The varying chemical characteristics of the different classes of compounds will necessitate several analyses, but it is possible to standardize the use of a limited number of protocols that rapidly captures the bulk of small molecules. A number of analytical approaches are currently available that can image a large number of metabolites, but they need to address the problems of co-eluting interference, and be able to accurately identify as many of the peaks as possible. The current status of promising analytical approaches and what is needed to forward these approaches will be the focus of this plan. Included are sample preparation needs, description of the advantages and disadvantages of a given analytical approach and how a combination of multiple approaches can circumvent limitations. Overall, the current-best protocols need to be modified for high throughput, while simultaneously developing the next generation of high-throughput protocols that are scalable and address difficult-to-measure metabolites. Classes of challenging metabolites include intermediate-sized molecules (1000-2500 Da), and charged molecules, such as phosphorylated compounds. In addition to determination of the steady-state concentrations of large numbers of metabolites, in many cases, it is the flux of these metabolites that will provide key insight in to which gene was perturbed.

The concentrations and fluxes of metabolites will need to be assessed in biochemical pathway inference models to probe pathway linkages. Recognizing that the greatest gains in functional genomics analysis will be derived from the integration of the different data streams, tools and approaches that combine the different classes of genomic data that are available, including DNA sequence data, mRNA expression profiles, protein profiles, and metabolite profiles, need to be developed.

SCIENTIFIC OBJECTIVES

Sampling and Sample Preparation

Given that many compounds are unstable, have very high turnover rates, or exhibit diurnal variation in concentration, etc., sampling is important. Sample extraction prior the chemical analysis must be adapted for each type of sample (e.g., extraction protocols for leaf tissue from *Populus* might be different from extraction protocols for developing xylem tissue). The metabolites represent many different classes of compounds, and therefore the chemical properties of the metabolites are highly variable. Depending on the extraction protocol, different classes of compounds show different extraction efficiency under specific conditions. It is unlikely that a single extraction procedure for plant tissues allows accurate quantification for all compounds, but the goal is to capture as many metabolites as possible.
Short-term goal: To establish standardized extraction protocols that are tailored to each tissue-type of *Populus* with high recovery and reproducibility. The extraction protocols will include addition of internal standards representing all major classes of compounds (e.g. carbohydrates, organic acids, steroids, amines, etc.) prior to extraction to increase accuracy and precision of the analysis.

Long-term goal: To establish methods for reproducible fractionation of extracts using solid phase extraction (SPE) columns or other methods. The goal of fractionation is to concentrate metabolites, permitting more of the extract to be analyzed by the gas chromatography (GC) - or liquid chromatography-mass spectrometry (LC-MS), and lessening the probability of saturating columns or MS-detectors.

Derivatization for GC-MS analysis

GC-MS analysis of extracts containing such varied metabolites as organic acids, sugars, sugar alcohols, amino acids and steroids is complicated. Many of the metabolites are not volatile and must be derivatized prior analysis by GC. Methoxymation in combination with trimethylsilylation (methoxy-TMS) is widely used as the main derivatization protocol. By first protecting the carbonyl group(s), the coupled derivatizations are more efficient (than silylation alone) for low molecular weight organic acids, but the relatively low temperature of the protocol may limit the derivatization of the more difficult to derivatize metabolites, including secondary carbon compounds that are typical of *Populus*. The organic acids that are most vulnerable to TMS derivatization alone can also be captured in other separations and analyses.

Short-term goal:

- Confirm the method(s) that minimizes sample preparatory time and maximizes spectral output (i.e., maximum metabolites observed) of the *Populus* species under investigation. The protocols must ensure the stability of the derivatized compounds and reproducibility of the data generated.

Long-term goal:

- Develop standardized extraction and derivatization protocols that are suitable for complete automation.

Analytical Techniques for Steady-State Metabolite Analyses

High-throughput GC-MS

Comparison of GC-MS techniques: The "oldest" and best-established coupling of methods to MS for metabolite analysis is GC-MS. The thermo-stable samples are vaporized and then ionized by either electron-impact (EI) or chemical-ionization (CI). For metabolomics analyses there are in practice two types of instruments used: 1) quadrupole (single stage MS and two stage MS/MS (ion trap)) and 2) time-of-flight (TOF) MS instruments. Both instruments have advantages and disadvantages, e.g. the quadrupole instruments have large dynamic range, are robust and easy to use. GC-TOF instruments have the capability to perform rapid spectral acquisition (up to 4-500 spectra/s over the full mass range), which results in possibility to speed up the analyses (high throughput), as narrow, short GC columns are used. The deconvolution of overlapping peaks is also greatly improved because of spectral continuity across a peak (no skewing of different masses). A disadvantage with GC-TOF instruments has been a reduced dynamic range (in practice) compared to quadrupole instruments, which can be a problem when analyzing complex samples with high variation in concentrations of compounds. GC-TOF MS is the approach of choice for high-throughput GC analyses, but as instrument manufacturers produce new quadrupole instruments with much higher scan rates, their ease-of-use may make them more versatile.

1D vs 2D GCxGC-MS

Rapid advancement in two-dimensional (2D) gas chromatography (GCxGC) makes it a powerful tool when coupled with high-speed TOF-MS for the deconvolution of metabolites that co-elute in traditional one-dimensional (1D) GC-MS. The GCxGC-TOF-MS approach couples columns of different polarity and operated at different temperatures to shift retention times of co-eluting metabolites. The peak capacity is approx. equal to the product of the separation capacities of the individual columns. The eluent from the first column is pulsed into a second column, generating an array of high-speed secondary chromatograms that can be detected by the high speed, high capacity time-array detector of TOF-MS. All of the eluent from the first dimension is subjected to separation in the second dimension, not just a single congested area of the chromatogram. The approach can be used for deconvolution to ensure the proper assignment of fragments to the metabolites being deconvolved, which can then be incorporated into the data extraction algorithms of the 1D analysis. As a deconvolution tool, it can be applied with the introduction of each novel heterogeneous matrix (e.g., a new poplar species/tissue). Given that the approach

requires the second column to function much faster than the first column, it is unlikely that GCxGC-TOF-MS will be deployed (in the near-term) as the standard data acquisition approach for high throughput profiling, until detectors with even more rapid acquisition rates become available.

Library Compilation of Mass Spectral Fragmentation Patterns

The effective deployment of GC-MS approaches for metabolite profiling must include the expansion of MS databases to identify unknowns for establishing data extraction and deconvolution strategies (metabolites quantified free from co-eluting interference). A greatly enlarged mass spectral database of EI and CI fragmentation patterns of TMS- and methoxime/TMS derivatives of metabolites will be required to identify as many of the large number of metabolites as possible. Up to 70% of the metabolites in a complex GC-MS analysis are typically unidentified. Although commercially available GC-MS databases have a large number of compounds, they have only a small proportion of the large number of low-molecular weight (<1200 Da) organic metabolites that need to have the fragmentation pattern of their methoxime/TMS derivatives characterized. *Populus* has a large number of unique phenolic compounds, including salicyl alcohol and higher-order conjugates with simple phenols (e.g., salicyl alcohol, catechol, 4-ethylphenol), phenolic acids (e.g., caffeic acid, benzoic acid, salicylic acid), and their glucosides. Compounds previously reported to be present in Salix sp. have a high likelihood of also being observed in *Populus* sp. Given that many plant species have some of metabolites in common, the data emerging from plant species, such as Arabidopsis and potato (Solanum), can be exploited to expand the metabolite database of mass spectra for metabolite profiling that will be an essential tool for phenotyping. Elucidation of many of the currently unidentified compounds is imperative and would provide an invaluable resource for metabolite profiling. Unidentified peaks can be subjected to more detailed characterization by subjecting the same samples to CI by using methane plasma to generate a milder ionization source, increasing the probability of observing the molecular ion ((M+1)+, (M-1)-). CI is of great benefit to identify the eluting peaks by confirming molecular weight and aiding in attempts to deconvolve overlapping peaks by simplifying the spectra (i.e., minimal fragments).

Short-term goals:

- Compilation of EI and CI mass spectral fragmentation patterns of *Populus* metabolites, including identification and localization of 500 of the most abundant metabolites with respect to key fragment assignments and retention time of their TMS and methoxime/TMS derivatives by GC-MS

Mid-term goal:

- Develop of data deconvolution and extraction algorithms/strategies to accurately quantify metabolite concentrations free from co-eluting interference

Long-term goal:

- Develop a centralized, web-based searchable repository of EI and CI mass spectral fragmentation patterns of TMS and methoxime/TMS derivatives of known *Populus* metabolites

LC-MS

In recent years, protocols for assessing metabolites using LC coupled to MS have been established with high analytical precision and sensitivity. Profiling methods based on this approach have been established for isoprenes, alkaloids, phenylpropanoids, glucosinolates, flavonoids, saponins and oxylipins. However, polar, non-volatile samples and LC conditions require special ionisation techniques:

ES. Electrospray ionization (ES) is the optimum method of ionizing the widest range of polar metabolites. Initially the sample is dissolved in a solvent where, to a certain extent, it will exist in an ionized form, e.g. [M-H] + ([M-H-]). In conventional ES the solution is then pumped through a thin capillary, which is raised to a high potential. Small charged droplets are sprayed from the ES capillary into a bath gas at atmospheric pressure and travel down a pressure and potential gradient towards an orifice in the MS high-vacuum system. As the droplets traverse this path they become desolvated and reduced in size until the point is reached that either an ion desorbs from a droplet or solvent is completely removed. The result of ES is a beam of ions, which are sampled by the mass spectrometer. ES is a concentration- rather than a mass-dependent process, and improved sensitivity is obtained for high-concentration low-volume samples, which has led to the development of low-flow-rate ES (e.g. nano-HPLC).

MS/MS. MS combined with ES was developed from single analyzer systems to very complex MS/MS couplings. Principle: In an MS/MS experiment a precursor ion is mass-selected by mass analyzer 1 (MS1) and focused into a collision region preceding a second mass analyzer (MS2). The process can be continued for several further steps (MSn). The mass analyzers are arranged in series either in space (sector, triple quadrupole and hybrid instruments) or in time (trapping instruments). Inert gas is generally introduced into the collision region and collisions occur between the precursor ion and inert gas atoms (molecules). In these collisions part of the precursor ion's translational energy can be converted into internal energy, and as a result of single or multiple collisions an unstable excited state is populated. Excited precursor ions decompose to product ions in a process termed “collision-induced dissociation” (CID). Product ions are mass-analyzed by MS2. There are also further alternative methods of precursor-ion dissociation that include surface-induced dissociation (SID), black-body infrared radiative dissociation (BIRD) and electron-capture-induced dissociation (ECD).

Hybrid MS/MS. While each type of MS/MS instrument has its own strengths and weaknesses, by combining analyzers in a hybrid conformation, an attempt was made to accentuate positive features while canonizing the negative ones. The hybrid MS/MS instruments that use a quadrupole as MS, and an orthogonally arranged reflectron TOF as MS (e.g. Q-TOF), are also commercially available and holds promise in metabolic profiling.

Direct Infusion (DI)-MS and MS/MS can be developed as a next-generation high-throughput screening tool for characterization of the large number of silent phenotypes, as an initial screen, without the use of time-demanding chromatographic resolution. A wide array of metabolites can be directly injected (infused) into a tandem MS that is operated with multiple ionization modes. Samples can be quickly (30 sec/analysis) subjected to both APCI and ES in the positive and negative ion modes to ionize different components within the heterogeneous matrix. The

analysis can especially target metabolites that are otherwise typically difficult to derivatize or analyze by GC-MS (for example), because of large size (500-1500 Da), too many reactive functional groups, and low volatility.

FTICR. Another direct infusion MS technology is Fourier transform ion cyclotron MS, wherein extracts are infused into MS instruments using soft ionization techniques to gain fingerprints of molecular ions present in an extract. FTICR instruments show promise in that they also provide exceptional high resolution (HR) and mass accuracy. The power of this technique relies on the presence of a mass analyzer capable of generating mass data that is sufficiently accurate for determination of definitive empirical formulae for several hundred ions.

Short-term goals:

- Establish standardized LC-MS analyses for metabolomics, progressing towards high throughput rates by developing direct infusion analyses (e.g., DI-nanoES, FTICR). Although LC-MS fragmentation patterns are minimal with respect to fragment generation, compilation of available ES spectra, including retention time and key fragment assignment (molecular ion in [M-H]+ ([M-H-] forms most likely) would promote metabolomic analyses by targeting those classes of compounds that are not readily analyzed by GC-MS approaches.

Mid-term goal:

- Analyses must move beyond simply detecting metabolites, but approaches and strategies developed in tandem to deconvolute the complex spectral output. Therefore, data analyses should also approach the throughput rates at which the analyses are conducted.

Long-term goal:

- Once the protocols are refined, a fully automated robotic sample preparation and handling systems can be developed for even greater sample throughput.

Pressure-assisted chemical electrophoresis-MS (PACE-ES-MS) is ideally suited for the efficient separation and analysis of highly polar, charged metabolites (e.g., polyphosphorylated compounds, adenylates (e.g., ATP, ADP), nucleotides (e.g., NADP, NAD, FAD), and coenzymes). The mass detection limit with a sheath-flow interface is ca. 10 to 100 fmol. Sensitivity can be increased by converting to a sheathless interface between the PACE and the MS ion trap. The approach will capture the many phosphorylated carbon intermediates, and can provide assessments of metabolite flux, when coupled with tracing stable isotope-labeled precursors and intermediates through pathways with MS and NMR analyses of isotopomers, as discussed below.

Short-term goals:

- Demonstrate utility of PACE-MS to profile phosphorylated carbon intermediates and coenzyme A (CoA)-bound intermediates, and trace stable isotope-labeled precursors and intermediates through pathways over time to determine metabolite flux rates.

Long-term goal:

- Increase throughput by the development and incorporation of chip technology and/or instruments with large numbers of capillaries. Once the technologies are refined, a fully automated robotic sample preparation and handling system can be incorporated for even greater sample throughput.

Transitioning from Steady-State Analysis of Metabolites to Metabolic Dynamics

One of the most exciting but daunting challenges in conducting global metabolic analyses are to create tools that facilitate the transition from quantifying static pools of metabolites to analyzing metabolite dynamics. Four approaches that are essential for analyzing metabolite dynamics are: 1. temporal quantification of metabolites in living tissues, 2. spatial analysis of metabolites in tissues, 3. experimental analysis of flux through one or more metabolic pathways, and 4. modeling of metabolic flux using bioinformatics tools.

In vivo NMR is a non-invasive, non-destructive technique for both qualitative and quantitative analysis of 13C- and 15N-labelled compounds, making it the procedure of choice to quantify metabolites in live, intact tissues, as well as to trace the metabolic fate of labeled compounds in real time. Recent studies indicate that it may also be feasible to modify existing in vivo NMR protocols to quantify metabolites by means of endogenous isotope levels. In vivo NMR technology has been used to study several plant species, including forest trees. However, there are a plethora of variables that affect the detection of high-resolution spectra, and there is substantial methods development to be done, including the lowering of detection limits, in order to proficiently examine trace regulatory metabolites in different poplar tissues.

It will also be important to investigate emerging instrumentation and technologies for in vivo spatial and temporal metabolite analyses. Novel imaging techniques, currently being developed within the medical bioengineering domain, such as in vivo positron emission tomography, hold considerable promise for quantifying metabolic dynamics in living cells, tissues, and organs.

Short-term goals:

- Develop standardized protocols for in vivo NMR analysis of different poplar tissues. Develop freely-available libraries of peak assignments.

Mid-term goals:

- Develop protocols for improved resolution and improved detection of metabolites using in vivo NMR.

Long-term goals:

- Exploit emerging imaging instrumentation for enhanced resolution real-time spatial and temporal metabolite analyses.

In addition to experimental protocols and instrumentation, bioinformatics tools are needed to model metabolic flux. These tools would either be selected from pre-existing tools, or created de novo. The tools will need to be parameterized for the pathways of interest, and then validated

with real data. Eventually, it will be advantageous to be able to overlay transcript and protein profiling data onto metabolic flux models.

Short-term goals:

- Assess the utility of various existing programs for modeling metabolic flux.

Mid-term goals:

- Develop new programs or customize existing programs for modeling metabolic flux. Create a database for publicly available metabolite flux data.

Long-term goals:

- Integrate metabolic flux modeling program with a poplar Pathway/Genome Database.

Data Analysis

Chemometrics and Visualization of Metabolite Data

Metabolomics projects generate large sets of data, and the accepted way of comparing large data sets is to use different multivariate tools, e.g. principal component analysis (PCA) or partial least squares projections to latent structures (PLS). PCA is an unsupervised method where no "a priori" knowledge of the class of samples is needed, and it is based on calculation of latent variables. The principle components are linear descriptions of the original descriptors, and are uncorrelated. The components also describe decreasing amount of data variance, i.e. P1>P2 and more. PCA will show the best representation of metabolic variation to be described in a limited number of dimensions. PLS is supervised method so that the class of a sample from an independent data set can be predicted on the basis of a series of models that are derived from the original data. This will help maximize the separation between classes, but also enable the data validation. Although the use multivariate statistical tools for evaluation of data are critical to handling all the generated data, it is also important develop tools for visualization of obtained data. The possibility to interpret differences between samples is dependent on describing metabolic differences in visually simple ways, relating relative differences with known metabolic pathways to pinpoint impacted steps (reactions).

Short-term goals:

- To establish a common strategy for evaluation of data by using multivariate statistical tools as PCA and PLS when appropriate. To establish tools to rapidly visualize metabolic differences.

Long-term goals:

- To establish a web-based (or software based) system for automatically compare large series of e.g. GC-MS and/or LC-MS data. A web-based tool where the identified metabolic differences observed can be visualized.

Systems Integrations Approach to Functional Genomics

While the focus of this first 5-year Poplar Science Plan is on tool and resource development, these tools and resources are being developed with an eye towards functional genomics, which is the projected goal of the second 5-year Plan. At the heart of functional genomics is delineating the function of genes identified via sequencing projects. For genes encoding enzymes, the most basic descriptors of function are enzymatic activity, substrate(s) and product(s). Thus, together with transcript profiling and protein analysis, global metabolite analysis represents one of the three cornerstones to delineating gene function.

It follows that integrating sequence data with transcriptomic, proteomic, and metabolomic data will be central to assigning function to genes. Bioinformatics tools such as Pathway/Genome Databases combine sophisticated database capabilities with gene ontology in order to permit integration of transcriptomic, proteomic, and metabolomic data with annotated whole genome sequence data. These tools often use graphical representations of metabolic networks as a visual, interactive interface for data integration and data mining. Thus, graphical, interactive metabolic networks form an important anchor for these integrative databases. The overall goal for this 5-year Science Plan is to create an interactive, queriable version of Pathway/Genome Database for Poplar.

Short-term goals:

- Create PoplaCyc, a Pathway/Genome Database containing only metabolic and biosynthetic pathways known (or suspected) to exist in poplar.

Mid-term goals:

- Add interactive functions to PoplaCyc to enable visualization of genomic, transcriptomic, proteomic, and/or metabolomic data within the graphical environment of PoplaCyc, and to allow for integration of datasets. Steps will include:
    - Annotate PoplaCyc with queriable *Populus* locus identifiers such as genomic sequence IDs and GenBank accession numbers. These locus identifiers will be linked to databases containing *Populus* gene function information, such as Gene Ontology.
    - Enable links between queriable *Populus* locus identifiers in PoplaCyc and *Populus* locus identifiers in microarray datasets, permitting gene expression data to be interactively overlaid onto PoplaCyc.
    - Enable links between metabolite identifiers in PoplaCyc and metabolite identifiers in metabolite analyses datasets.

Long-term goals:

- Add previously unknown metabolites and undetermined metabolic pathways to PoplaCyc.
- Add cellular activities to PoplaCyc, such as signal transduction, transmembrane molecular transport, and molecular trafficking.
- Link PoplaCyc to public databases of gene expression and metabolite datasets.

Strategies and Potential Future Applications

PoplaCyc could be created using available tools such as Pathologic, which was used to create MetaCyc, a compendium of metabolic networks, and the species-specific BioCyc databases, such as AraCyc (Arabidopsis). AraCyc should serve as a model for PoplaCyc. It would be important to work with the developers of Pathologic and AraCyc. It will also be necessary to attract the necessary bioinformatics specialists to achieve these goals in a timely manner.

PoplaCyc has the potential to evolve into a pivotal, higher-order bioinformatics tool that allows for easy integration of other diverse metabolic and genomic data to create a "virtual interactive tree." Such a tool will allow for in silico dissection of complex events such as responses to environmental perturbation.

Protein Characterization & Proteomics

Panel Members: **Christophe Plomion**, Fredrik Sterky, Cetin Yuceer and Malcolm Campbell.

BACKGROUND AND SCOPE

While the Populus genome project provides an overview of the genes in poplar trees, it is becoming increasingly clear that this is only a very fragmentary beginning of understanding their role and function. The old paradigm of one gene for one protein is now considered to be an oversimplification, as it does not account for the vast potential of alternative splicing and post-transcriptional modification to give rise to a variety of gene products from each gene. Current knowledge suggests that each gene is more likely to give rise to 6-8 proteins per gene in eukaryotes. Consequently, genome information alone is insufficient to define all potential proteins. Furthermore, while transcript abundance is a very useful indicator of gene expression, it does not provide a complete picture of the range of proteins that are produced. Transcript profiling does not provide information related to protein turnover, post-translational modifications such as signal peptide cleavage, phosphorylation or glycosylation, sub-cellular localization of proteins, or the complex interactions between proteins. Unfortunately, these processes cannot be deduced from microarray or nucleic acid-based methodologies. Thus, while transcript profiling provides a snapshot of the potential of the genome to give rise to diverse cellular functions, it provides only one layer of information. Given that the transcriptome data is only indicative of the cell's potential and does not reflect the actual state in a given cell at a given time, the concept of "proteome," (PROTEin complement expressed by a genOME), has emerged to provide complementary and critical information by revealing the regulation, activities, quantities and interaction of every protein in the cell. Consequently, proteomics is now considered as a priority by many universities and research institutes and is starting to be widely applied to the model plant Arabidopsis and other important crop species. However, forest tree proteomics in general and more specifically poplar proteomics still remains largely embryonic, in spite of the fact that a large collection of expressed sequence tags (ESTs) is available and a 6X genome coverage for poplar is now available. The Populus proteomics effort will result in discoveries about the cell's protein machinery that could yield important applications in forestry. Such knowledge could lead to an understanding of the molecular basis of soft and hardwood formation, reproductive development, juvenility and maturity, and biotic and abiotic stress responses.

Proteomics has now branched into five specific disciplines: protein profiling, delineation of protein-protein interactions, post-translational modifications, protein localization, and structural analysis.

Protein Profiling:

The standard approach to protein profiling has capitalised on the power of combining high-resolution Two-Dimensional Polyacrylamide Gel Electrophoresis and Mass Spectrometry (2D-PAGE-MS). Proteins are resolved on a 2D-PAGE and identified using MS. This technique enables differential display proteomics for comparison of protein abundance in diverse contexts. Protein solubility and detection issues can still be improved. Recently, alternative technologies

(MudPIT, ICAT, and Protein Chips) have been described to resolve, identify and sometimes quantify many of the expressed proteins in complex protein mixtures. The following are the potential technologies that can be developed and made available to the researchers working on Populus genomics.

2D-PAGE-MS: is the most widely used method for protein separation, quantification and identification. Recent improvements in 2D-PAGE, like new IPG strips and more sensitive staining techniques, have improved resolution and reproducibility. In large number of samples, thousands of protein spots can be individually displayed, quantified, compared, and identified by using dedicated software. Mass spectrometry has proven to be effective in (i) identifying proteins at the femtomoles to picomoles levels, (ii) evaluating post-translational modifications, and studying macromolecular complexes. In addition, dramatic gains in analytical sensitivity already allow proteomic analyses to be carried out on small number of cells, which largely eliminate the need for molecular amplification strategies. The accessibility to mass spectrometers is therefore considered as a critical point in post-genomic projects.

Further improvements and refinements should be sought in 2D-PAGE-MS. 2D-PAGE detects the most abundant and soluble proteins, resulting in poor separation and visualization of total protein content of a cell. Low abundance proteins like transcription factors and protein kinases are not detected on standard 2D-gel protein maps. To partially overcome this problem, proteins can be separated into sub-fractions according to cell type or sub-cellular compartments, such as plasma membrane, chloroplast, or mitochondria. It is often difficult to dissolve nuclear and hydrophobic membrane associated proteins, but detergents in combination with thiourea seem promising for the analysis of these proteins by 2D-PAGE.

Multidimensional Protein Identification Technology (MudPIT): Multidimensional liquid chromatography has recently been developed to resolve low abundance proteins, proteins with extreme pI and Mr, and integral membrane proteins. However, it is not suitable for the quantitative detection of differences in protein abundance of cells in different states.

Isotope-Coded Affinity Tags (ICAT): This recent technology enables a high throughput and direct comparison of the relative abundance measurements of the same peptide ion from two sample populations after affinity-based purification. In a mixture, the two populations of peptides are individually labeled with two different ICAT reagents, one being eight mass units heavier than the other. Fractions are analyzed by MS following the mixture being passed through affinity media. The difference in protein abundance is then estimated directly by the difference in peak height of the two corresponding peptide ions. ICAT is only used for proteins containing cysteine and not suitable for the complex analysis of many samples or the study of the effect of several factors.

Delineation of Protein-Protein Interactions:

Since many cellular processes involve multiprotein complexes, the identification and analysis of protein interaction networks are the keys to understand how the ensemble of expressed proteins (proteome) is organized into functional units. Several approaches to study protein interactions include co-purification or direct purification of complexes by bi-dimensional chromatography. The classical yeast two-hybrid system results in high rate of false positive and is associated with

the inability to define protein complexes that involve many disparate P-P interactions. New techniques such as TAP-TAG (Cellzome), FLAG-TAG (MDS proteomics), CHH-TAG (CSHL), and GFP-TAG-TAG (Bordeaux University) coupled with mass spectrometry allow the characterization of multiprotein complexes in a large-scale approach. The study of the complexome is one of the main objectives of far more advanced functional genomics projects such as Human, Yeast and Arabidopsis. We therefore consider highly redundant to address this level during the first phase of the poplar genome project. Rather, one should use knowledge (protein-protein interaction map) and reagents (e.g., antibodies) from these advanced projects to address biological questions that are either unique or important to trees, by focusing our investigations at the individual protein level.

Post-translational Modifications:
Proteins may undergo some form of modification following translation. Modifications like proteolytic cleavage, phosphorylation, methylation, or glycosylation affect both the properties of proteins and then interactions with other molecules. For example, glycosylation plays critical roles in protein sorting and receptor binding. These modifications result in mass changes and cannot be determined from the genomic sequence or mRNA expression data. Thus, they can only be addressed at the protein level. Mass spectrometry and immobilized metal affinity chromatography have been used to study such modifications.

Protein Localisation
Antibody Proteomics: Antibodies are excellent tools for proteomics with their ability to uniquely bind to protein epitopes. Monoclonal antibodies are very expensive, but recently, a less expensive and streamlined protocol was developed to produce monospecific polyclonal antibodies (www.hpr.se). The antibodies can be used for tissue and subcellular localization of proteins as well as for protein expression profiling and analysis of protein size variations. In essence, a nucleic acid sequence corresponding to a small part of the protein (100-150 aa) is designed, amplified by RT-PCR, and cloned into an expression vector. A fusion protein is produced containing an albumin binding part and a His-tag for affinity purification. After immunisations, the polyclonal antibodies are purified using the same protein used for immunisation to increase the specificity of the antibodies. The technology is based on a combination of established protocols that has been optimised for higher throughput. The fragments used are big enough to generate a good immune response but small enough to enable high success rate in *E. coli* production. Also, by using only parts of proteins, the uniqueness of fragments can be ensured and trans-membrane regions (and other problematic regions of proteins) can be avoided. The technology is under development but has been successfully applied on human genes in combination with tissue microarrays. The main bottleneck is the cost of immunisations.

Subcellular proteome analysis: The identification of proteins recruited to fulfill the specific function of subcellular compartments constitutes a step forward to proteome analysis. Protocols have been optimized for the extraction and solubilisation of proteins after sub-fractionation according to cell type and subcellular compartments, e.g. cell wall, plasma membrane, vacuolar membrane, endoplasmic reticulum, Golgi apparatus, mitochondrion and chloroplasts. It is foreseen that the study of "subproteomes" will contribute to the recovery of much more proteins than usually revealed by classical proteomic experiments.

Localisation tagging: By determining the sub-cellular location of proteins and its pattern of expression, important clues to their functions can be gained. The classic way of studying protein localization is to use enzyme reporters, such as β-glucuronidase (GUS). However, this technique suffers from low resolution and other techniques have become more attractive. One alternative is to use antigenic tags (for example T7 or HA epitopes) or specific antibodies. Another convenient tool is fusion of the target protein to the green fluorescent protein (GFP) or some of its spectral variants. This technique provides a sensitive tool to study sub-cellular localization in real time. N- or C-terminal fusions of GFP may in some cases yield localization artefacts, which can be avoided using the FTFLP (Fluorescent Tagging of Full-length Proteins) approach. In FTFLP, the fluorescent protein is incorporated internally and the tagged protein is expressed under native regulatory conditions.

Structural Analysis:
Biochemical and cellular investigations help identify the function of proteins and their interactions, such as the interaction between a ligand and its receptor. The three-dimensional structure of the interaction could give detailed information on the active site and amino acid residues that are involved in the interaction. This would lead to more specific functional studies like site-directed mutagenesis to gain insights into the way proteins work. X-ray diffraction and NMR spectroscopy are the two main technologies for determining the three-dimensional structures of proteins.

SCIENTIFIC OBJECTIVES

Short-term goals:

- Define specific biological questions where proteomics will generate valuable pieces of information, rather than accumulate extensive data.
- Use a rational approach to prioritise which "subproteomes" will be characterised.
- Develop/adopt protocols to improve sensitivity, resolution, reliability, throughput, and information of 2D-PAGE-MS for diverse *Populus* tissues.
- Optimise protocols for immunostaining of multiple *Populus* tissues and apply antibody proteomics for a limited number of key genes. Develop a community-wide antibody inventory (a physical collection that could be covered by Genetic Resources panel).
- Develop a common web-based database to store, query, track *Populus* proteome data:
    - Define a minimum community standard for 2D PAGE–MS representation to facilitate data comparison, exchange and verification.
    - Develop a public repository center for mass spectrometry data where the many levels of MS data are stored and are available for data mining (raw data, peak lists, peptide identification, and protein identification). Such repository should also include the description of the experiment to allow users to query across experiments to observe concomitant changes across studies. This repository center should of course have access to all Populus ESTs and to the genomic sequence.
    - Provide researchers with direct links from each annotated gene to all available data from the proteome studies, including selected external databases.

Mid-term goals:

- Profile *Populus* proteomes at various developmental stages and in response to agriculturally important biotic and abiotic signals.

Long-term goals:

- Identify biochemical functions of unique proteins
- Delineate protein-protein interactions and develop an understanding of post-translational modifications under various biological/developmental and environmental conditions for targeted proteins.

Strategies to achieve the goals

- Overall there is a profound lack in our community of appropriately trained scientists in proteomics. Developing a network (linked to the IPGC) will help to capitalize on proteomic expertise and instrumentation existing in few forest tree biology laboratories. This network will (i) review the existing facilities by surveying the proteomics capacity in the community (see annex), (ii) organize the cross-laboratory facilities, and (iii) try to stimulate research in poplar proteomics.
- Establish a publicly available virtual proteomics laboratory, from which current and new technologies will be available to achieve the proposed goals.
- Develop training programs in proteomics.
- Develop a state-of-the-art information system for storing and mining proteomic data in a harmonized manner in connection with the Informatics, Annotation & Database Development group. In its simplest form, proteomic analysis should play a role in genome annotation by providing evidence that a particular gene is expressed in vivo and under which circumstances and by providing the number of protein isoforms encoded by a particular gene, and the molecular nature of the isoforms (e.g. phosphorylation).
- Define the genetic features of proteins (such as polymorphisms, mode of transmission) in connection with the Physical Mapping & Marker Development group.
- Compare proteome, transcriptome and metabolome data in connection with the Gene Expression/Microarrays and Metabolic Characterization & Metabolomics groups.

Study the genetic variability of qualitative and quantitative protein variants

- Map the expressed genome (using conventional genetic mapping strategy applied for proteins)
- Evaluate the modifications of protein expression in respect to genetic factors (PQL mapping)
- Determine the extent to which quantitative phenotypic variance is mirrored by quantitative variation of proteins (QTL/PQL coincidence)

Study the variability of proteome expression in physiological studies

- 'Differential display proteomics' between organs, developmental or physiological stages and contrasted abiotic or biotic conditions will be used to identify clusters of proteins for which the expression is idiotypic for a specific state. Technologies such as 2D-PAGE, ICAT and Antibody proteomics could be used.

Establish predictive mathematical descriptions of biological systems

- By combining DNA sequence, mRNA profiles, protein expression and metabolite concentration as well as information about dynamic spatio-temporal changes in these molecules.

SUMMARY

Tools for high-throughput protein profiling are available within the IPGC community. They now need to be organized to avoid redundancy. We suggest that a consistent proteomic approach should be taken by the different participants of the international consortium and that reference maps of *Populus* tissue should be established on the World Wide Web for constant updates and comparisons of different studies. A relational database for plant proteomic has recently been created, in which members of the IPGC have contributed (Dumazet et al. 2004 *Proteomics*, in press). This tool could be the starting point for sharing proteomic data in poplar.

Gene Expression & Microarrays

Panel Members: **Peter Nilsson**, Wout Boerjan, Jörg Bohlmann, Amy Brunner, Steve DiFazio and John MacKay

BACKGROUND AND SCOPE

DNA microarrays have during the last few years made an impressive impact in the research area of functional genomics. The technology of spotted cDNA microarrays was pioneered at Stanford University, in the latter part of the nineties, with the creation of the first whole genome array for yeast. This was a major breakthrough in microarray technology for transcript profiling, when the enormous possibilities were realized with simultaneous analysis of whole transcriptomes. This led to a dramatic increase in microarray research.

Another factor that affected the breakthrough of microarrays was the parallel establishment of photolithographically in situ synthesized oligonucleotides and methods to analyze hybridization signals from them, developed by Affymetrix, Inc. As more and more sequences became available through the Human Genome Project and other large-scale sequencing efforts, it became possible to design short, gene specific oligonucleotides to be used as immobilized probes. However, this latter methodology has not yet reached the same broad distribution in academia, mainly because the costs were initially very high.

Finally, microarrays for transcript profiling based on spotted long oligos (50-70-mers) have recently been introduced. The primary advantage of this approach is the reduced effort needed to prepare the probes for immobilization and the possibilities to design oligos for certain purposes such as the analysis of splice variants or coverage of closely related species.

The majority of microarrays for organisms that are not completely genome sequenced have been based on amplified cDNA. These are usually produced in the context of EST-sequencing projects, with the creation of unigene-sets through sequence assembly, preparation of clones by cultivation, plasmid preparation, PCR, PCR-purification, and transfer to a suitable spotting plate. For some of the major model organisms, i.e. those that have been completely sequenced, it is now also possible to design oligos through bioinformatic efforts and synthesize or buy them for spotting.

Another alternative is to design gene-specific PCR primers and individually amplify fragments of all genes, as has been done in the CATMA project for Arabidopsis.

CURRENT STATE OF THE FIELD

UPSC/KTH, Sweden
A large scale Tree Functional Genomics program has been run in Sweden by the Umeå Plant Science Centre in collaboration with KTH, Stockholm. The first *Populus* arrays that were produced were tissue-specific arrays with approximately 3000 wood and leaf clones, respectively. The wood chip has been utilized for a number of studies such as full-scale analysis of very thin wood-forming tissue sections (Hertzberg et al, 2001, PNAS). The very limited

amount of starting material for target preparation led to the development of a methodology for amplification of minute samples. It is based on capture of short 3'-tags of fragmented cDNA, which enables non-biased PCR amplification and labeling (Hertzberg et al, 2001, Plant J). Subsequently a unigene collection based on 33.000 EST sequences was assembled, rendering 13.500 clones which constitute the POP1 set. More than 2000 POP1 arrays have been produced at KTH since Nov. 2001 and they are now the basis for many different experiments for several groups both in Umeå and in Stockholm, as well as several external collaborators. The next generation, denoted POP2, will be based on 100.000 ESTs and a unigene set consisting of 25.000 clones. It is now being prepared and is estimated to be ready to spot in mid 2003.

A research project led by Chung-Jui Tsai at Michigan Technological University has produced 11,000 ESTs from various tissues of aspen (shoot tips, young leaves, young stems, root tips) that have been deposited in GenBank. 6K microarrays are being produced for gene expression analysis as part of a Michigan Life Sciences Corridor-funded functional genomics research project to investigate various aspects of growth and metabolism. The 6K arrays consist of >5,000 unique sequences with some degree of redundancy due to contig-forming overlapping clones.

Pilot studies using spotted oligos are also underway. Helge Küster, University of Bielefeld, Germany: 2400 oligos (70-mers from Operon) selected for their association with the symbiosis between Poplar and Mycorrhiza. The Forest Biotechnology Group at Oregon State University is using a 230 oligo (70-mer) array of selected transcription factors and related regulatory genes to study changes in gene expression associated with maturation and flowering. Oak Ridge National Laboratory is producing design software and an array of 1500 70mer oligos to test the ability to differentiate closely related members of gene families while still retaining functionality across the genus.

SCIENTIFIC OBJECTIVES

A whole-genome microarray would be an invaluable resource for the poplar community, and the most expeditious way to accomplish this is through international collaboration. The aim is to establish a global collection of DNA probes for spotting onto microarrays, covering all genes in *Populus*. Importantly, it should be possible to utilize them for the majority of the different *Populus* species and subtypes being used for research. One of the major objectives is also to establish a strong and active microarray group within the IPGC, working for standardization and development of analysis tools through a common database. It is also important that the availability of spotted microarrays is not restricted to a limited number of research groups. This could be achieved by assigning some resource centers responsibility for the production and distribution of microarrays.

Short-Term Goals:

The initial aim will be to establish a consensus strategy toward the common goal of a complete whole-genome array. This will be done within all research groups with interests in *Populus* microarrays. The aim is to enable a coherent effort to combine existing and planned microarray-related *Populus* projects, both in terms of sharing knowledge and experiences and not the least, physical resources. Limited local efforts for different species and genotypes could be combined

through exchange of DNA probes and/or sequences. The most important strategic issue to decide upon is what kinds of array-immobilized probes to use. The information that will be achieved from the ongoing genome sequencing and annotation will definitely be a necessary knowledge base for any attempts to generate a whole genome array. Even if an existing large-scale EST-sequencing effort has been accomplished to create a cDNA microarray covering substantial parts of the genome or rather transcriptome, it is hard to foresee that cDNA will be the ultimate choice in the end. Instead it is easy to see the advantages of spotted long oligos, not least in terms of possibilities for complete coverage and also for the possibilities to design oligos to conserved regions, enabling a broad usage for the whole *Populus* genus. Problems will of course appear for the genomes which are not sequenced, but different EST sequences could possibly aid in the design process. Furthermore, spotted longmer oligos are advantageous for the flexibility, ease of preparation and exchange of information regarding oligo sequences.

Mid-Term Goals:
The availability of the complete and annotated genomic sequence will enable the initialization of the bioinformatics process to design gene-specific oligos. The aim is to collect all publicly available *Populus* sequence information and also encourage different research groups in the consortium to provide sequences which are not yet publicly available. A coherent bioinformatics effort is needed.

A relative high financial investment is initially needed to purchase complete oligo sets, even though the actual production costs for the arrays are drastically reduced. However, it is possible that a deal could be reached with a company to do a large-scale synthesis and sell small batches to individual groups at a reasonable price.

Long-Term Goals:
The long-term goals are to have a solid global organization for all *Populus* related microarray activities, centralized around the production and utilization of a common whole genome array.
A complete awareness of the availability of the commonly achieved whole genome spotted oligo arrays will provide extremely good opportunities for the whole *Populus* research community to perform many interesting experiments. An ultimate aim is also to integrate all results in a common *Populus* expression database, which will enable important cross-experiment data mining.

Future Applications

A whole-genome array will be an invaluable tool in poplar functional genomics for the foreseeable future. The array will continue to provide a comprehensive view of transcriptome differences among tissue types and along developmental gradients, allow in-depth characterization of responses to environmental perturbations, and facilitate the deconvolution of complex regulatory pathways through the analysis of mutant lines. In addition, the array will undoubtedly be used in novel ways such as to derive traits for Quantitative Trait Locus analyses, or to generate character traits for phylogenetic analyses. As the costs of array production, labeling, and hybridization continue to fall, the range of experiments will grow, providing an increasingly detailed view of the functional genomics of this important model organism.

Informatics, Annotation & Database Development

Panel Members: **Francis Martin**, Jan Karlsson, Dan Weems, Loren Hauser, Natalie Pavy and Pierre Rouzé.

BACKGROUND AND SCOPE

Aims of this panel include the formation of publicly accessible, readily updateable, globally linked intraspecific and interspecific genomics databases. Three main tasks include:

1. Development of bioinformatic resources
2. Annotation of the *Populus* genome, including
    a. Development/tuning of software & tools
    b. Structural (syntactic) annotation : modeling of genes for the whole genome
    c. Functional annotation: give functional attributes to every gene and/or gene product.
3. Development and curation of databases

CURRENT STATE OF THE FIELD

Physical and Financial Resources

A number of groups and individual laboratories exist in Europe and North America that are involved in gene prediction and genome annotation, database curation, BLAST servers, gene profiling, functional classification of unigene sets, comparative genomics, studies of genome duplication and gene families, EST annotation and clustering, SAGE analysis of transcriptional regulatory elements and microarray data analysis .

The Plant Systems Biology Group at Ghent University is a partner in PLANeT, an EU-founded initiative coordinated by Klaus Mayer (MIPS) which aims at providing plant scientists access to plant genomics data, knowledge and resources collected in the different partner countries, and to share curation tasks and expertise. INRA-Bordeaux is coordinating the EU network EVOLTREE (involving up to 200 scientists from a dozen of European countries), aiming at investigating tree biodiversity and involving genomics as the first component. Poplar is one of the three species chosen in this program, and helping its annotation one of the milestone of the proposal. The University of Tennessee, University of Minnesota and JGI NSF-funded *Populus* genome portal project will develop a *Populus* EST database and associated search and analysis tools; integrate genetic mapping and QTL data; develop a *Populus* microarray resource database and associated search and analysis tools; and curate the *Populus* genome annotations.

SCIENTIFIC OBJECTIVES

Short-term goals:

- Generate a data-set of non-redundant full length cDNAs and ESTs to build a relevant database. Up to 2000 full-length cDNAs and 200K ESTs are currently available.

- Check (mostly through automatic routines) all poplar BACs for these documented genes, as well as evolutionary conserved genes. Annotate automatically & check manually all these annotations. Build the relevant training sets from this annotation set.
- Validate the existing comparative annotation programs for adequacy to comparison of poplar and Arabidopsis genomes.
- Depending on the results, fine-tune one or several of them, and possibly develop additional in-house capabilities
- Enter in database specific genome features to be filtered out (or annotated separately) in the annotation process (e.g., repeats, rRNAs, transposons, etc.)
- Fine-tune and optimize the *ab initio* components of EUGENE_POP and/or GRAILEXP (Markov models for exon, introns, UTRs and intergenic, splice site predictors, and translation start sites)
- Develop syntactic (structural) annotation of the poplar genome, BAC-wise using EUGENE and POP/GRAIL-EXP
- Provide comparative evaluation of syntactic annotation. Depending on sequencing status, provide a provisional complete genome annotation (with or without functional annotation, see below)
- Build routines to collect comparative functional annotation.
- Routine first functional annotation through database matches (SwissProt, Gene Ontology and Interpro)
- Enter Ontology consortium and validate for poplar

Mid-term goals:

- Compare POP/GRAIL-EXP and EUGENE gene models (and any other modeler that IPGC scientists want to use). This comparison will be on going for a number of years. To be discussed at the yearly meetings.
- After an agreed upon time and as more full length cDNAs and assembled ESTs become available, remodel the entire genome with retrained POP/GRAIL-EXP and EUGENE gene models.
- Finalize annotation and distribution of the workload. To be discussed at the yearly meetings
- Train a number of poplar biologists on how to do annotation since they will be the ones who will primarily use and edit the final poplar database.
- Development of the Poplar Genome Anatomy Project (PGAP) aiming to determine the gene expression profiles of poplar tissues/cells, leading eventually to improved detection and diagnosis for the economically-relevant traits. The PGAP will provide comprehensive genomic data, including expressed sequence tags (ESTs), gene expression patterns, single nucleotide polymorphisms (SNPs), cluster assemblies, and cytogenetic information, together with informatics tools to query and analyze the data and information on methods and resources for reagents developed by the project.

Long-term goals:

- Build a visualization tool to understand architectural structure of gene expression in trees (linking gene annotation, Gene Ontology, microarray gene expression profilings).

Strategies and potential future applications:

Gene prediction & annotation in poplar will be largely using tools (POP/GRAIL-EXP /EUGENE) that have been developed for other genomes (e.g., Arabidopsis). Nevertheless several points have to be noticed:

- Incomplete coverage of *Populus* sequencing will have a negative influence on the performance of *ab initio* gene finding.
- Due to the relatively low numbers of ESTs (200 K) and FL cDNAs (1500) gene finding will gain only marginally from data from the expressed genome (contrary to human or even Arabidopsis & Rice).

SUMMARY

There will be soon several plant genomes entirely sequenced, or with large amounts of sequence data available (Arabidopsis, rice, Medicago, maize). Gene prediction in poplar should use comparative genomics to a large extent. This approach has been promoted recently for the human genome with more closely related organisms (human/mouse). We will need to fine-tune or re-develop these existing tools to cope with anomalies in the gene models.