



PostgreSQL Scientific Application - Case example

PostgreSQL Genomic Databases

Sébastien Clément

sclement@cfri.forestry.ca

Natural Resources Canada

Presented at the PostgreSQL Conference 2009 in Japan

November 20th



Natural Resources
Canada

Canadian Forest
Service

Ressources naturelles
Canada

Service canadien
des forêts

Foreword

« What is this guy doing here ? »

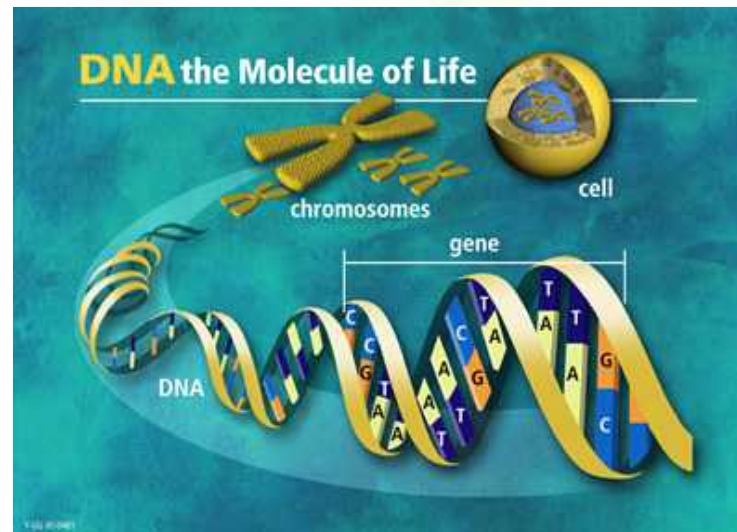


SKEPTICAL CAT

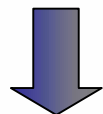
« Can PostgreSQL handle scientific databases ? »

What is genomics and why bother ?

Genomics: « *The study of the entire genome (all genes) of a species* »



3 000 000 000
~23 000

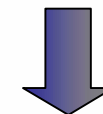


- Health and disease
- Heredity
- etc.

Genome size
Number of genes



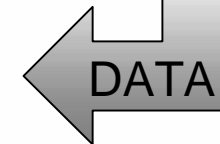
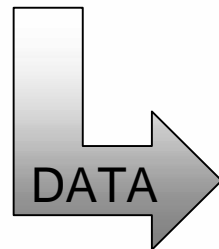
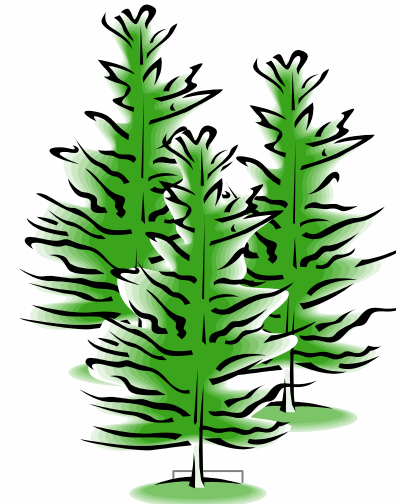
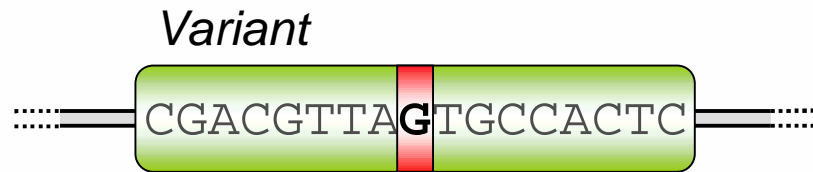
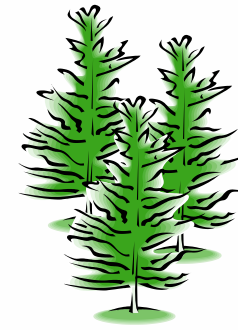
390 000 000
~53 000



- Genetic improvement

Why study *TREE* genomics?

Cellulose gene



Why is a genomic DB essential?

A single gene...

CGACGTTAATGCCACTC

Species

Name

Sequence

Variations

Size

Functions

Chromosome pos.

```

ACCCCTGGGATGATTACGGTCTTTCTGGGAAACAGTGGAGCCCATGATATT
GATGGGAAACGAGCTTCCCTCGCCCTTGTCTATCTTTTCTCGTGAGAAGAGACC
TGGATATCAGCCACCACAAAGAACTGCTGCCATGAACGCCCTGGTTCCGAG
TCTCTTACGTTTCTAACAAATGCCCCATTTCATCTTTGAACCTGGACTGTGAT
CCCTAAAGAGGCGCTGGCGGAGGGCAATGTGTTTTCTCAT
GGATCGGAGGAGGAACTGCTGCTACGTTTCAGTTCCCTCAGAGAT
TCCGATGGTATTGATCGCAATGATCGGATATGCAAAACAGAAACACTGTTTTCT
TTCCGACATCAACATGAAGGCTCTGGATGGCATCCAGGGTCCAGTGTATGT
AGGGACTGGATGCGTTTTCAACAGCCAACTCTGTATGGGTATGATCCTC
CAGTTTTCCAGAAAGAACTAAGATGACCTGGGATTGCTGGCCATCATGG
TGTTGCTGTTTTGGCGGTTCTCCCAAGAAACAAAGAAATCATCCAAAAA
GCTCTTTGGTAGGAAGAAATCTTCCAAAGCCAACTGAAATAACAGCTCCCA
TCTTCAGCCTGGAAGAAATGAAAGAAAGCACTTCAAGGTTACGAAGAACAT
GAGAAATCATGGTTAATGTCGCAGAAAGCACTTTGAAAAAGAGATTTGGTCA
CTCTCCTGCTTTTATTACATCCACGCTTATGGAAAAAGGTTGGTGTTCAG
ACTCTGTAAACTCACCTTCTCTGATAAAGCAAGCCATTTCACGTCATAAGC
TCTCTGATAAAGCAAGCCATTTCACGTCATAAGC
GAGAAAACTGAAATCGGGCAAGCAGATCGGATGGATATA
CAGAGGATATCCTGACAGCTTTTCAAAATGCACTGCCGTG
CTTGGCGGATCGGTATACTGCATGCCCACTAGCGCCGGCATTTCAAAGGATCT
GCTCCTATTAATTTGTGAGATCGTTTGCACCAGCTTCTTCGTTGGCGCTT
CGGTTCCATCGAAATTTTCATGACTAGACATTGCCCACTCTCGGTATGCC
ATCCAGGAAACCTAAAAATGGCTTCAAAAGATTGCCCTATGTCAAATACAATC
GTTTACCCGTTTCACTCAATCCCGCTCATCGCTTACTGCACGCTCCCGCG
CATATGCTTCTCACAGGAAATTCATCACTCCAAGCTTACAAAGCCTTTG
CTAGCGTCTGGTTCATGGGCTCTTTCTTATCCATCATAGCCACTGGTGT
CTACAGCTGAGATGGAGCGGCACTGACCATCCAAGAGCTTTCGAGAAACGA
ACACTTCTGGGTCACTGGAGCGGCTCTCTGCTCACTTTTTTGGCTGTTTTTC
AAGCCCTGCTCAAGGTGTTGGCGGCTCTGGACACCAAAATTTTACAGTCACA
GCCAAGCGTAGCCATGAAGAAAGCAAAATTTGGGGAGTTATACATGTTTCAA
ATGGACGACCCTGCTCATTTCCCTCCCACTACCTTGCTGATAAATAAATCTGG
TGACCGTGGTAGCAGGTGTTTCCGCTGCAATAAATAAATAACTACCAATCA
TGGGACCGCTCTTTGGAAAGCTCTTCTTCCGCTTCTGGGTGATATTGCA
TCTGTATCCCTTCTTCAAGGCTCTTCTGGCAAGGCCAGAACCGGAACGCCCA
CTATCGCTGCTACTGTGGTCTATTCTGCTGGCCCTCCATCTTCTCGCTGTG
TGGGTGAGGATCGACCCATTCCTGCCCAAGTTCGAAGCTCCCATTTCTTCA
GCAATGTCGCGTCCGACTGTTAACTACCTCCTCTCTTCTGATCATTTTATGA
ATTCTATTTCATTGATTCAGATTTTGTAAATGATTCGAAGGATAAAACTG
AAAGCATTTCTATTATAGTTTGTGAATGCTATTTCATTCCGATGCAAAATA
TTGTAATCAACCGAAGCATAAAAACACAGGCAATTCGACTCGTACTGTAAGA
TTCCGATGCAATCTTTAAAAAATTTGCAATAATGAAGACTTGGCC
GCTGATGTTA
    
```

more...
(phew!)

Interaction with
other genes

Similarity with
other species

Cell wall metabolism

Cell structure

Glycolysis

**...how about thousands of genes...
...for thousands of species?**

Public genomics DBs

Genbank



NCBI GenBank Overview

PubMed Entrez BLAST GMM Backs Taxonomy Structure

Search Entrez

NCBI Home

NCBI Site Map

Submit to GenBank

Submit an update

Search GenBank

GenBank and RefSeq a comparison

BLAST

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2009 Jan 26; *Database Issue* 37: D25-30). There are approximately 106,531,156,756 bases in 108,431,690 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,057 sequence records in the WGS division as of August 2009.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank [record](#) may be viewed for a *Saccharomyces cerevisiae* gene.

In The News: 2009 H1N1 Flu Virus (Swine Flu)

The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the [NCBI Flu Resource](#).

<http://www.ncbi.nlm.nih.gov/Genbank/>

UniProt



EMBL-EBI UniProt

All Databases Entrez Text Home

Databases Tools EMBL Outlets Training Mobility About Us Help Site Index

UniProt Home

UniProt Website

About UniProt

Search Tools

UniView

Submissions

Downloads

Documentation

Publications

People

Help

UniProt - Welcome to UniProt

The mission of UniProt is to provide the scientific community with a comprehensive, high quality and freely accessible resource of protein sequence and functional information. UniProt is composed of four components, each optimized for different uses. The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-references. It consists of two sections: **UniProtKB/Swiss-Prot** which is manually annotated and reviewed, and **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed. The **UniProt Reference Clusters (UniRef)** databases provide clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences. The **UniProt Archive (UniParc)** is a comprehensive repository, used to keep track of sequences and their identifiers. The **UniProt Metagenomic and Environmental Sequences (UniMES)** database is a repository specifically developed for metagenomic and environmental data.

The sequences and information in UniProt is accessible via text search, BLAST, [UniProt Search](#), and [FTP](#).

<http://www.ebi.ac.uk/uniprot/>

TAIR



tair Home Help Contact About Us Login/Registrar

Search Browse Tools Stocks Portals Download Submit News

The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABCR are fully integrated into TAIR.

Breaking News
[Subscribe to news feed](#)

New AraCyc (PMN) release [October 15, 2009]
[The Plant Metabolic Network](#) is pleased to announce new releases of the metabolic pathway databases AraCyc 6.0 and PlantCyc 3.0, plus the inaugural release of PoplarCyc 1.0.

Stock Donation Made Easy [July 24, 2009]
ABRC and NASC are currently seeking donations of characterized mutant lines. We have a recruitment sheet!

TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation.

<http://www.arabidopsis.org/>

General views

- Summary
- Genes and contigs
- Gene family
- Summary family
- Genes
- Contigs
- Classified
- Expression
- Primers
- PCRs
- Sequencing
- SNPs
- Logout

Gene family 4CL

Classified contig10660 (dir8)

Gene family: [4CL](#)

SNP 08Pg10660a

TCCAGCCGGACGA**Y**GTCGTGGCGTTGCCTTATTCATCCGGAACAACGGGG
 CTCCCCAAGGGCGTAATGCTAACGCACAAGGGCCTGGTGTCCAGCGTTGC
 GCAGCAGGTCGACGGTGA **[C/G]**AACCCCAATCTGTATTTCCATTCCGGAG
 GACGTGATTCTGTGTGTCTTGCCTCTCTTCCACATCTACTCTCTCAATTC
 GGTTCTCCTCTGCGCGCTCAGAGCCGGGGCCGCGACACTGATTATGCAGA
 AATTCAACCTGACGAC**R**TGTCTGGA**S**CTGATCCAGAGATATAAGGTTACA
 GTTGCCCCAATTGTGCCCCCGAT**Y**

Gene pos.: Exon - codant (CDS)

Mutation type (CDS): CDS/non-synonyme

Chips (2)

Chip	To be send on chip ?	Mapping chip success ?	Nat. pop. chip success ?
PGLM1	<input checked="" type="checkbox"/>	✓Polymorphic	✓Polymorphic
WD1	<input checked="" type="checkbox"/>	✓Polymorphic	✓Polymorphic

[GQ0011.BR F15 \(1.2\)](#)

[33 \(dir8\)](#), [01428 \(dir8\)](#), [01718 \(dir8\)](#)

[435 \(dir8\)](#), [13148 \(dir8\)](#), [12582 \(dir8\)](#),

ATATTA ATATGCATAT
 ACTTTC CATAIGATAC
 CGTTGC ACCGATTAT
 >>>>>> >>>
 CATCTG GAAATGATA
 TTCATT TTGCTGCAAG
 GACGGC GACTTGGGAA
 GTAGAA AACCACTGC
 TCTCGG ATTTCTTAC

TA 4-coumarate--CoA ligase (EC 6.2.1.12)

4CL-4	13148 (dir8)	<input type="checkbox"/>	4
4CL-5	12582 (dir8)	<input type="checkbox"/>	4
4CL-6	14172 (dir8)	<input type="checkbox"/>	4

SNPs (5):

Name	Position	♀	♂	Pool	Mapping parents priority.	Mapping pool priority.	Illumina score
08Pg10660a	120F	✓	✗	✓	1		0.860
08Pg10660b	264F	✓	✗	✓	2		0.595
08Pg10660c	335R	✓	✗	✓	2		0.554
08Pg10660e	320F	✗	✓	✗	1		0.841
08Pg10660f	530F	✗	✓	✓	2		0.678

[Add a contig](#)

[Edit](#) | [Gene families](#)

TreeSNPs overview (cont'd)

Lab plate



Plate view

Results

PCR results for primer plates 008-F and 008-R

New PCR plate for primer plates 008-F and 008-R
 Enter PCR results for PCR plates 008-F and 008-R

PCR-functional primer pairs: 57
 Paralogous primer pairs (i.e. haploid sequencing failed): 9

008-F 008-R	01	02	03
A	8207-F 8207-R >= 620 bp TD-58 (TDA) : OK (800 bp)	8247-F 8247-R >= 607 bp TD-58 (TDA) : !	8347-F 8347-R >= 599 bp TD-58 (TDA) : OK (600)
B	8984-F 8984-R >= 553 bp TD-58 (TDA) : OK (800 bp)	9016-F 9016-R >= 670 bp TD-58 (TDA) : OK (900 bp)	9039-F 9039-R >= 623 bp TD-58 (TDA) : OK (400)
C	9572-F 9572-R >= 610 bp TD-58 (TDA) : OK (800 bp)	9629-F 9629-R >= 691 bp TD-58 (TDA) : OK (700 bp)	9673-F 9673-R >= 526 bp TD-58 (TDA) : OK (600)
D	10143-F 10143-R -- 500 bp	10175-F 10175-R -- 500 bp	10211-F 10211-R -- 200 bp

Primer plates 008-F, 008-R

Name: 008-F, 008-R

[Download in CSV \(plate format\)](#)

008-F 008-R	01	02	03	04	05	06	07	08	09	10	11	12
A	8207-F 8207-R	8247-F 8247-R	8347-F 8347-R	8489-F 8489-R	8558-F 8558-R	8564-F 8564-R	8581-F 8581-R	8613-F 8613-R	8620-F 8620-R	8815-F 8815-R	8822-F 8822-R	8944-F 8944-R
B	8984-F 8984-R	9016-F 9016-R	9039-F 9039-R	9057-F 9057-R	9227-F 9227-R	9361-F 9361-R	9394-F 9394-R	9489-F 9489-R	9514-F 9514-R	9527-F 9527-R	9559-F 9559-R	9569-F 9569-R
C	9572-F 9572-R	9629-F 9629-R	9673-F 9673-R	9679-F 9679-R	9749-F 9749-R	9802-F 9802-R	9823-F 9823-R	9861-F 9861-R	9928-F 9928-R	9935-F 9935-R	9938-F 9938-R	9955-F 9955-R
D	10143-F 10143-R	10175-F 10175-R	10211-F 10211-R	10224-F 10224-R	10238-F 10238-R	10290-F 10290-R	10295-F 10295-R	10392-F 10392-R	10424-F 10424-R	10480-F 10480-R	10561-F 10561-R	10805-F1 10805-R1
E	9531-F1 9531-R1	10707-F1 10707-R1	10076-F1 10076-R1	7749-F1 7749-R1	10282-F1 10282-R1	10282-F2 10282-R2	10282-F3 10282-R3	5745-F1 5745-R1	5227-F1 5227-R1	5227-F2 5227-R2	10672-F1 10672-R1	10672-F2 10672-R2
F	10052-F1 10052-R1	10375-F2 10375-R2	9622-F1 9622-R1	9622-F2 9622-R2	15972-F1 15972-R1	6279-F1 6279-R1	8776-F1 8776-R1	8776-F2 8776-R2	7410-F1 7410-R1	5586-F1 5586-R1	1070-F2 1070-R2	8806-F1 8806-R1
G	6487-F2 6487-R2	1782-F2 1782-R2	11144-F2 11144-R2	11283-F2 11283-R2	4241-F3 4241-R3	3163-F1 3163-R1	15158-F1 15158-R1	10968-F1 10968-R1	10096-F1 10096-R1	3064-F1 3064-R1	6696-F1 6696-R1	10215-F1 10215-R1
H	9293-F1 9293-R1	15589-F1 15589-R1	12531-F1 12531-R1	5440-F1 5440-R1	6053-F1 6053-R1	11855-F1 11855-R1	12982-F1 12982-R1	10388-F1 10388-R1	10388-F2 10388-R2	10581-F1 10581-R1	10581-F2 10581-R2	10213-F1 10213-R1

[Add/remove primers from plates 008-F, 008-R](#)

Primer plates | [PCR results for primer plates 008-F, 008-R](#)

TreeSNPs overview (cont'd)

Example of calculations (views):

Summary by gene family

Gene families: 129

Notes:

000NC: other unidentified contigs, to be classified

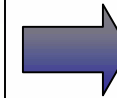
In red: Success rates below 50% (PCR, sequencing) or 30% (global success) .

Gene family	Contigs	Unique genes	Analyzed regions	Primers									SNPs		
				Total	Synth.	PCR			Sequencing			Global success	Total	Per unique gene	Per region
						Tested	OK	%	Tested	OK	%				
000NC	4374	0	2125	1198	1182	1141	717	63%	627	423	67%	42%	3043	-	1.4
4CL	18	15	6	12	12	12	12	100%	8	6	75%	75%	18	1.2	3.0
ACT	8	0	8	10	10	10	4	40%	2	2	100%	40%	5	-	0.6
AGO	7	0	7	12	12	12	10	83%	10	10	100%	83%	27	-	3.9
AGP	53	17	24	44	22	44	38	86%	34	20	59%	51%	75	4.4	3.1
AP	2	0	2	2	2	2	2	100%	2	2	100%	100%	8	-	4.0
AP2	58	22	44	63	39	63	57	90%	53	51	96%	87%	99	4.5	2.3
ARF	2	0	1	2	2	2	2	100%	2	2	100%	100%	5	-	5.0
ASF	1	1	1	2	2	2	2	100%	2	2	100%	100%	3	3.0	3.0
AUX-IAA	55	23	31	56	50	53	47	89%	45	41	91%	81%	121	5.3	3.9
BGAL	13	6	15	28	28	28	16	57%	14	10	71%	41%	47	7.8	3.1
bHLH	18	0	17	33	33	32	32	100%	26	24	92%	92%	70	-	4.1
bZIP	58	17	44	68	42	67	47	70%	38	36	95%	66%	79	4.6	1.8
BZR	1	0	1	2	2	2	2	100%	2	2	100%	100%	3	-	3.0

TreeSNPs overview (cont'd)

TreeSNPs download page & demo version

<http://treesnps-pub.arborea.ulaval.ca:3000/download>



Arborea TreeSNPs
Download and Installation

Overview

Downloads

- [TreeSNPs](#)
- [User manual](#)
- [Quick tutorial](#)
- [JFRC 2008 Poster](#)
- [Procedures](#)
- [Installation guide \(French\)](#)

DEMO version

- [Try now!](#)

Note: use guest as username and password

TreeSNPs was developed to store all data relevant to SNP of the *Arborea* white spruce genomics project.

It can be used to store, in logical order:

1. cDNA sequences
2. List of candidate genes, selected from (1)
3. List of primers developed on candidate genes (2)
4. PCR results using (3)
5. Sequencing results on (4)
6. Discovered SNPs on (5)

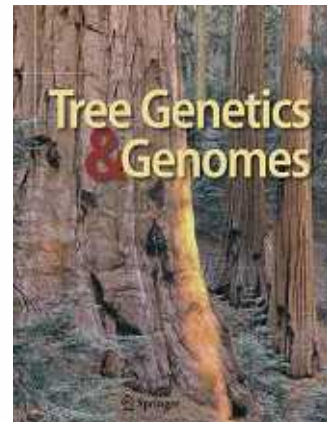
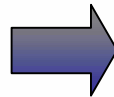
Given this data model, the database can be used for a variety of those steps.

TreeSNPs was designed to have a secure, concurrent, web-researchers and staff to a central database. Additionally to for data summarizing, analysis and export in a variety of for

SNP ID	Accession	Gene	Chr	Position	RefSeq	Gene	Accession	Gene
000001	U00001	U00001	1	1000000	U00001	U00001	U00001	U00001
000002	U00002	U00002	1	1000001	U00002	U00002	U00002	U00002
000003	U00003	U00003	1	1000002	U00003	U00003	U00003	U00003
000004	U00004	U00004	1	1000003	U00004	U00004	U00004	U00004
000005	U00005	U00005	1	1000004	U00005	U00005	U00005	U00005
000006	U00006	U00006	1	1000005	U00006	U00006	U00006	U00006
000007	U00007	U00007	1	1000006	U00007	U00007	U00007	U00007
000008	U00008	U00008	1	1000007	U00008	U00008	U00008	U00008
000009	U00009	U00009	1	1000008	U00009	U00009	U00009	U00009
000010	U00010	U00010	1	1000009	U00010	U00010	U00010	U00010

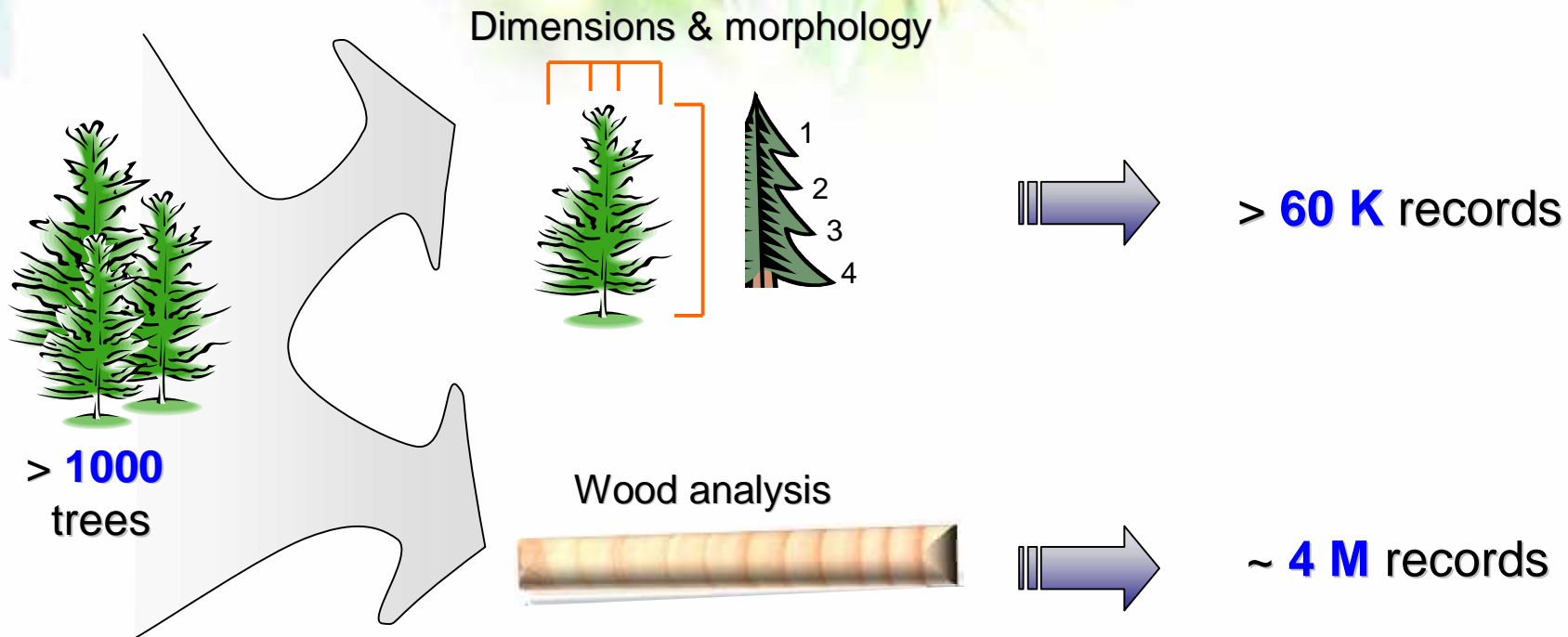
Adopted by U. of Alberta's (Canada) Laboratory on Mountain Pine Beetle

A paper to appear soon in



PhenoTree overview

What data is stored ?



Other data:

- Geographical locations
- Tree pedigree

PhenoTree overview (cont'd)

Wood analysis properties table

0 4.00 8.00 12.00 16.00 20.00 24.00 28.00 32.00 36.00 Radius (mm)

read every 25 µm

Pith

Bark

Fibre dimensions
Cells count

942 trees
×
~2100 reads/tree
= **1.98 M reads!**

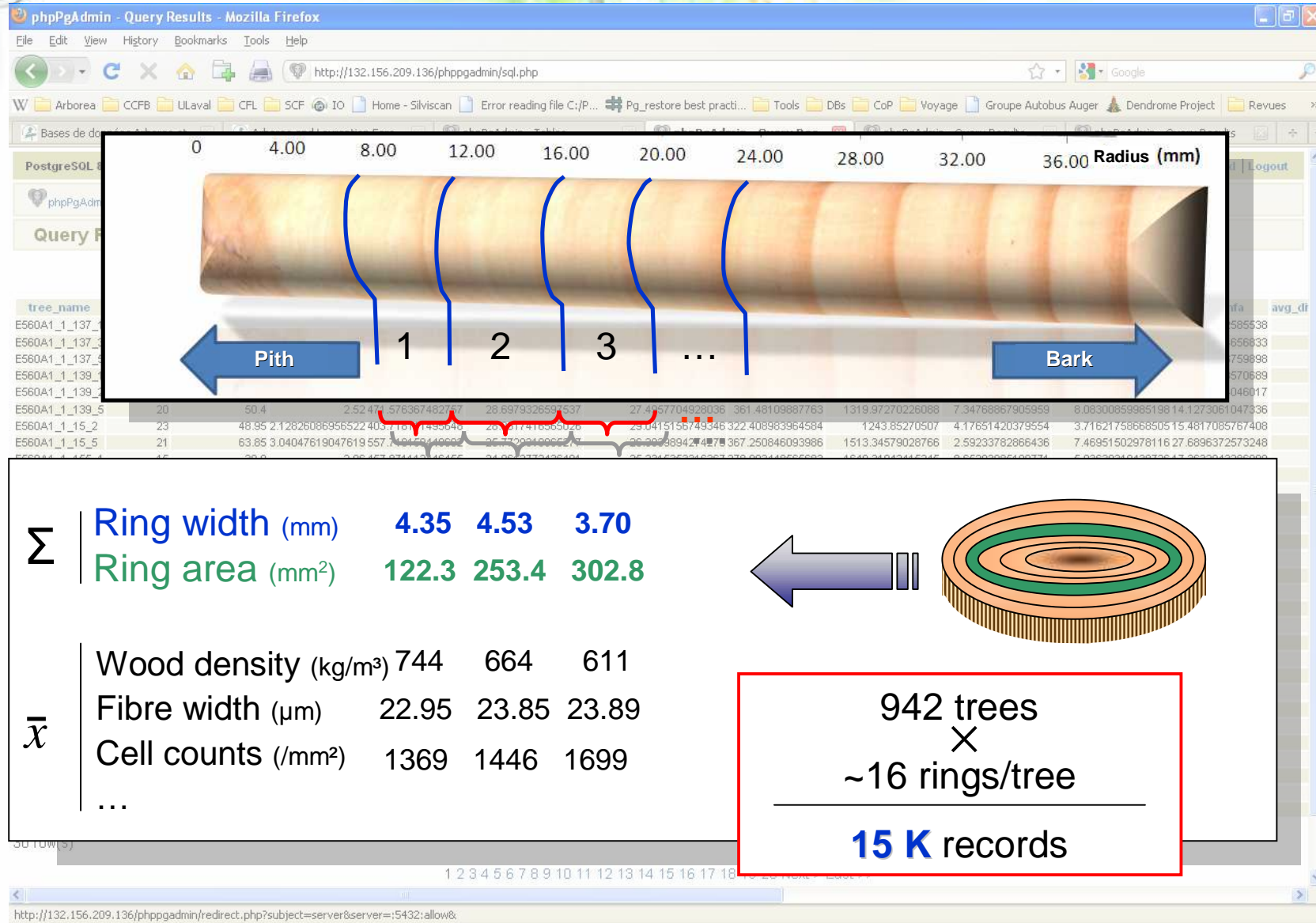
Wood density
etc.

Edit	Delete	E560A1_1_4_4	4	0.1	297	25.6	33.2	295	1165	22.6	39	34	0.143	3.8	1.52	463	2
Edit	Delete	E560A1_1_4_4	5	0.125	309	26.6	35.6	292	1057	22.6	39	34	0.143	3.8	1.65	426	2
Edit	Delete	E560A1_1_4_4	6	0.16	318	24.7	35.4	278	1142	22.6	39	34	0.143	3.8	1.63	432	2
Edit	Delete	E560A1_1_4_4	7	0.175	326	24.5	35	280	1165	22.6	39	34	0.143	3.8	1.66	426	2
Edit	Delete	E560A1_1_4_4	8	0.2	345	25.9	35.3	314	1096	22.6	39	34	0.143	3.8	1.82	389	2
Edit	Delete	E560A1_1_4_4	9	0.225	349	26.1	36	318	1096	22.6	39	34	0.143	3.8	1.85	384	2
Edit	Delete	E560A1_1_4_4	10	0.25	352	25.7	36	322	1096	22.6	39	34	0.143	3.8	1.88	379	2
Edit	Delete	E560A1_1_4_4	11	0.275	352	26.3	36	326	1096	22.6	39	34	0.143	3.8	1.91	374	2
Edit	Delete	E560A1_1_4_4	12	0.3	402	25.6	36	330	1096	22.6	39	34	0.143	3.8	2.15	334	2
Edit	Delete	E560A1_1_4_4	13	0.325	398	26.2	36	334	1096	22.6	39	34	0.143	3.8	2.18	331	2
Edit	Delete	E560A1_1_4_4	14	0.35	398	26.4	36	338	1096	22.6	39	34	0.143	3.8	2.28	318	2
Edit	Delete	E560A1_1_4_4	15	0.375	408	25.9	36	342	1096	22.6	39	34	0.143	3.8	2.34	310	2
Edit	Delete	E560A1_1_4_4	16	0.4	422	25.2	36	346	1096	22.6	39	34	0.143	3.8	2.4	303	2
Edit	Delete	E560A1_1_4_4	17	0.425	438	25.7	36	350	1096	22.6	39	34	0.143	3.8	2.5	293	2
Edit	Delete	E560A1_1_4_4	18	0.45	449	25.7	36	354	1096	22.6	39	34	0.143	3.8	2.68	274	2
Edit	Delete	E560A1_1_4_4	19	0.475	465	25.2	36	358	1096	22.6	39	34	0.143	3.8	2.75	267	2
Edit	Delete	E560A1_1_4_4	20	0.5	490	24.3	36	362	1096	22.6	39	34	0.143	3.8	2.72	269	2
Edit	Delete	E560A1_1_4_4	21	0.525	507	25.4	36	366	1096	22.8	39.2	34	0.143	3.8	2.67	275	2
Edit	Delete	E560A1_1_4_4	22	0.55	513	26	33.4	445	1152	22.8	39.2	34	0.143	3.8	2.72	270	2
Edit	Delete	E560A1_1_4_4	23	0.575	503	26.2	33.7	445	1132	22.8	39.2	34	0.143	3.8	2.83	260	2
Edit	Delete	E560A1_1_4_4	24	0.6	506	25.2	33.4	426	1189	22.8	39.2	34	0.143	3.8	2.85	278	2
Edit	Delete	E560A1_1_4_4	25	0.625	521	24.5	33.9	433	1204	22.8	39.2	34	0.143	3.8	2.83	260	2
Edit	Delete	E560A1_1_4_4	26	0.65	526	25.3	34.4	459	1146	22.8	39.2	34	0.143	3.8	2.75	268	2
Edit	Delete	E560A1_1_4_4	27	0.675	516	25.2	33.8	441	1176	22.9	39.3	34	0.143	3.8	2.65	278	2
Edit	Delete	E560A1_1_4_4	28	0.7	517	24.4	32.4	408	1267	22.9	39.3	34	0.143	3.8	2.63	280	2
Edit	Delete	E560A1_1_4_4	29	0.725	523	24	31.7	398	1314	22.9	39.3	31.8	0.182	7.3	2.63	280	2

30 row(s)

PhenoTree overview (cont'd)

Example of calculations 1 (SQL views): Growth ring averages



PhenoTree overview (cont'd)

Example of calculations 2 (SQL views): **crosstab function**

**Logical,
but not very useful...**

tree_name	year	height
E560A1_1_105_2	1986	220
E560A1_1_105_2	1992	450
E560A1_1_105_2	1997	640
E560A1_1_105_2	2005	940
E560A1_1_106_3	1986	230
E560A1_1_106_3	1992	510
E560A1_1_106_3	1997	670
E560A1_1_106_3	2005	1070
E560A1_1_124_1	1986	160
E560A1_1_124_1	1992	410
E560A1_1_124_1	1997	620
E560A1_1_124_1	2005	840

crosstab

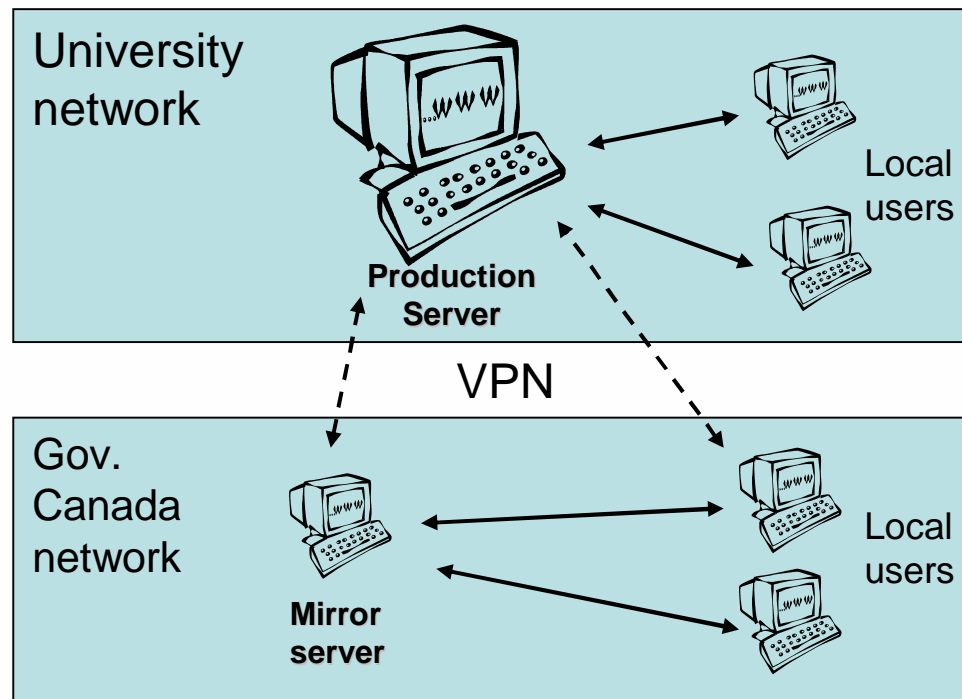


**...this is it what
end-users want**

tree_name	height_1986	height_1992	height_1997	height_2004	height_2005
E560A1_1_105_2	220	450	640	NULL	940
E560A1_1_106_3	230	510	670	NULL	1070
E560A1_1_124_1	160	410	620	NULL	840

Systems and user base

- Formerly Access projects (2006-7)
- Migrated to PostgreSQL 8.3 under Fedora (2007-8)
- Migrated back to Windows (2009)



- Around 20 scientific users (Universities, Federal Government)

PostgreSQL and Windows – can it really work ?

Task automation with DOS

- Limited functionality

Solution ?

Windows Task Manager

Cygun

Unix/bash scripts

PostgreSQL

Script examples:

- Start Rails server (DOS)
- Backups (DOS)*
- Backup files cleaner (bash)
- VPN connexion to production server (DOS)
- Mirror synchronizing (bash, DOS)
- Database version comparison (bash)
- Users & privileges report (bash)

*Thanks: Greg Smith (http://wiki.postgresql.org/wiki/Automated_Backup_on_Windows)

Developing databases for the scientific community

Suggestions:

- Have a user-based approach
 - 1. Know/answer the user's needs
 - 2. Limit technical jargon
 - 3. Think 'usability'

Aknowledgements

People

Jean Beaulieu – Lab director

Joël Fillon – Ruby on Rails interface designer

Jean-Philippe Dionne – Rails secure access programming

Jean Bousquet – Collaborator

All end users, particularly:

Sylvie Blais, Stéphanie Beauseigle, Marie Deslauriers,
Pier-Luc Poulin, Patrick Lenz

Organizations

Arborea Forest Genomics (<http://www.arborea.ulaval.ca/>)

Canadian Forest Service, Natural Resources Canada

Genome Québec

Genome Canada

Done ?

```
SELECT tmp_rings.tree_s_b_f_t, tmp_rings.experiment_id, tmp_rings.ring_pith_to_bark, bad_rings.ring_pith_to_bark IS NULL AS is_valid, tmp_rings.start_index,
tmp_rings.start_position, tmp_rings.end_index, tmp_rings.end_position, tmp_rings.width, tmp_rings.area, tmp_rings.avg_density, tmp_rings.avg_radial_diameter,
tmp_rings.avg_tangential_diameter, tmp_rings.avg_coarseness, tmp_rings.avg_cell_population, tmp_rings.avg_ray_angle, tmp_rings.avg_isopycnic_angle, tmp_rings.avg_mfa,
tmp_rings.avg_diffraction_intensity_cv, tmp_rings.avg_moe, tmp_rings.avg_wall_thickness, tmp_rings.avg_specific_surface, tmp_rings.avg_crystallite_width,
tmp_rings.avg_crystallinity, early_woods.start_index AS ew_start_index, early_woods.start_position AS ew_start_position, early_woods.end_index AS ew_end_index,
early_woods.end_position AS ew_end_position, early_woods.width AS ew_width, early_woods.width / tmp_rings.width AS ew_percentage, early_woods.area AS ew_area,
early_woods.avg_density AS ew_avg_density, early_woods.avg_radial_diameter AS ew_avg_radial_diameter, early_woods.avg_tangential_diameter AS
ew_avg_tangential_diameter, early_woods.avg_coarseness AS ew_avg_coarseness, early_woods.avg_cell_population AS ew_avg_cell_population, early_woods.avg_ray_angle
AS ew_avg_ray_angle, early_woods.avg_isopycnic_angle AS ew_avg_isopycnic_angle, early_woods.avg_mfa AS ew_avg_mfa, early_woods.avg_diffraction_intensity_cv AS
ew_avg_diffraction_intensity_cv, early_woods.avg_moe AS ew_avg_moe, early_woods.avg_wall_thickness AS ew_avg_wall_thickness, early_woods.avg_specific_surface AS
ew_avg_specific_surface, early_woods.avg_crystallite_width AS ew_avg_crystallite_width, early_woods.avg_crystallinity AS ew_avg_crystallinity, late_woods.start_index AS
lw_start_index, late_woods.start_position AS lw_start_position, late_woods.end_index AS lw_end_index, late_woods.end_position AS lw_end_position, late_woods.width AS
lw_width, late_woods.width / tmp_rings.width AS lw_percentage, late_woods.area AS lw_area, late_woods.avg_density AS lw_avg_density, late_woods.avg_radial_diameter
AS lw_avg_radial_diameter, late_woods.avg_tangential_diameter AS lw_avg_tangential_diameter, late_woods.avg_coarseness AS lw_avg_coarseness,
late_woods.avg_cell_population AS lw_avg_cell_population, late_woods.avg_ray_angle AS lw_avg_ray_angle, late_woods.avg_isopycnic_angle AS lw_avg_isopycnic_angle,
late_woods.avg_mfa AS lw_avg_mfa, late_woods.avg_diffraction_intensity_cv AS lw_avg_diffraction_intensity_cv, late_woods.avg_moe AS lw_avg_moe,
late_woods.avg_wall_thickness AS lw_avg_wall_thickness, late_woods.avg_specific_surface AS lw_avg_specific_surface, late_woods.avg_crystallite_width AS
lw_avg_crystallite_width, late_woods.avg_crystallinity AS lw_avg_crystallinity
FROM tmp_rings
NATURAL LEFT JOIN bad_rings
JOIN tmp_woods early_woods USING (tree_s_b_f_t, experiment_id, ring_pith_to_bark)
JOIN tmp_woods late_woods USING (tree_s_b_f_t, experiment_id, ring_pith_to_bark)
WHERE early_woods.transition = 1 AND late_woods.transition = 2
ORDER BY tmp_rings.tree_s_b_f_t, tmp_rings.experiment_id, tmp_rings.ring_pith_to_bark;
```

