PAPACLIMA: MARKER ASSISTED SELECTION FOR POTATO GERMPLASM ADAPTED TO BIOTIC AND ABIOTIC STRESSES CAUSED BY GLOBAL CLIMATE CHANGE

Partner: NEIKER

Dr. Enrique Ritter	(Scientist-in-charge)	NEIKER
Dr Leire Barandalla	(Scientist)	NEIKER
MSc Alba Alvarez	(Scientist)	NEIKER
Dr. Jose I. Ruiz de G	(Scientist)	NEIKER

Specific Objective 2: Detection of useful candidate genes (CG) for abiotic and biotic stresses and Analysis of the allelic variation in these CG

A 2.1 Library construction and RNA-Seq

Period: First 6 Months

In order to obtain appropriate RNA populations first a small bio assay was set up:

One supposed tolerant potato cultivar, Rudolph, and one susceptible cultivar, Atlantic, was chosen for the experiment.

Tubers from these cultivars were put under heat stress (46-48ºC), under cold stress (8-10ºC) and under drought stress without irrigation until in all treatments severe stress symptoms were observed. Leaf samples were taken from each treatment and from the unstressed controls and frozen at -80ºC until further processing.

Eight cDNA libraries were constructed from these materials, which are shown in Table 2.1. The libraries consisted of adapter-ligated and barcoded restriction fragments of ds-cDNA.

The following methodology was applied:

Each sample was powdered without thawing in a mortar using liquid nitrogen. Total RNA was extracted from each sample using the NORGEN Plant RNA Extraction Kit, following the instructions of the manufacturer. Then poly-A+ RNA was obtained from total RNA using 5' biotinylated oligo (dT) primer bound to paramagnetic beads coated with streptavidin (Dyanabeads M-280 Streptavidin, DYNAL, Oslo, Norway). First and second strand cDNAs were synthesized according to Sambrook et al. (1989). Double-stranded cDNA (ds-cDNA) was digested with AseI and TaqI (NEB Biolabs Inc, New Brunswick, NE, USA). followed by ligation of AseI and TaqI adapters with T4 DNA ligase (Invitrogen Inc, Barcelona, Spain).

Illumina Primers containing sequences complementary to the adapters and sample specific barcodes (Index sequences) were used in two rounds of PCR amplifications as described by Bachem et al (1998). Figure 2.1 shows the scheme for library construction and Table 2.2 shows the adapter and primer sequences used for this purpose.

After size selection and purification, samples were pooled with approximately equal concentrations and – after quality control using a BIOANALYZER the pool was sent for sequencing on an ILLUMINA MySeq sequencing platform (StarSeq GmbH, Germany).

The results and sequence analyses are pending.

Period: Month 7 to 16

However, during the sequencing process no cluster were formed and the first results were disappointing, probably due to the presence of the large amounts of Taq-Taq restriction fragments (RF).
Therefore, the experiment was repeated using this time the varieties Soprano and Kondor as susceptible and tolerant cultivars, respectively, and processed as described above.
In addition, a modification of the molecular procedure was introduced in order to enrich for Ase-Taq fragments and to eliminate the Taq-Taq RF in an intermediare step. For this purpose the adapter-ligated fragments were amplified with a linear PCR using only a biotinylated ASE adapter primer for 5 cycles. Then the enriched Ase-Taq and few Ase-Ase amplicication products were caoptured using streptavidin coated magnetic beads, while the Taq-Taq fragments were washed off.
After separation from the beads, control PCRs using only the Ase, the Taq and both adapter primers showed in gel that the procedure has been efficient, Afterwards the captured amplicfication products were processed as described above and send for sequencing using 2x250 bp MySeq. This time sequencing was very successful!

A total number of over **5.6 million** “read through” sequences were obtained from our digested cDNA libraries.
Table 2.2.1a (Annex Period Month 7 to 16) shows the characteristics of the experiment including the obtained sequence numbers in each case.

Sequences were processed by in-house developed Software (**Radlib**). This Software was basically available from other R&D projects, but had to be adapted to handle Illumina MySeq data.

Unlike ION Torrent/Proton data, MySeq sequences are delivered separated as R1 and R2 reads, and per sample, defined by a combination of I5 and I7 indexes, but also in Unix format, so that a conversion is required.
Due to the fixed bp length of the reads smaller RF sequences run into the adapter and have to be trimmed by removing adapter sequences. The larger sequences of the R1 and R2 reads can be merged by sing for example Flash.exe (https://ccb.jhu.edu/software/FLASH/).
Afterwards the sequences of all samples are combined.

Although ASPAM (used to analyse CG amplicons; see below)) and RADLIB are similar in post processing, the fundamental difference is that in ASPAM the CG sequences are known, but a priori unknown in RAD sequencing.
Therefore RADLIB basically blasts each sequence against all sequences and extracts packages of homologous sequences, each one representing an “RF-CG” since it is supposed to target transcripts. In this way a list of “CGs” are obtained which are further processed as in ASPAM (see below).
However, the whole process, particularly the extraction of RF packages, lasts several weeks, the computer running day and night, although the Software has already be optimized to shorten the processing time (which initially was calculated as several months).

Although the process of extracting of reference sequences is not yet ready due to the reasons mentioned above, we have processed the first set of **8420 Reference sequences**, particularly in view of this pending report and in order to test the whole procedure. The extractions and analyses will be completed in the near future.

Table 2.2.1b (Annex Period Month 7 to 16) **shows the Reference Sequences of RAD markers in PATLIB (BIOEDIT;** partial view)**.** The sequence lengths range from 50 to over 250 bp.

Table 2.2.1c (Annex Period Month 7 to 16) shows the Distribution of Reads of RAD markers over GT. A large degree of variation can be observed.

Table 2.2.1d presents the obtained polymorphisms in the RF-CG markers (partial view).
From the 8420 reference sequences, for **2395 RF** no consistent patterns were obtained, mainly due to low read numbers. A total of **1996 RF** were monomorphic. The other **4029 RF-CG** revealed between **2** and **5** patterns in the four samples A which were defined by 1 to 29 SNPs.

Table 2.2.1e Contains the concrete Allele Sequences, SNP and Pattern Definitions of all RAD markers from PATLib (partial view).

Table 2.2.1f (Annex Period Month 7 to 16) contains the CG Allele composition of the 8 potato tetraploid samples (partial view).

Annotations for the RF-CG were produced for the RF markers using Mercator Software.
Fig. 2.2.1g (Annex Period Month 7 to 16) shows the functional Distribution of the Annotations for RF-CG markers. As can be seen in this picture over 50% of the RF could be assigned to specific functions (42.4% unassigned), targeting a broad range of functions. Of particular interest are the **4.18%** of RAD markers which are assigned to stress functions.

Table 2.2.1h presents a partial view of the concrete Annotations obtained for the over 8000 RF-CG markers. For **4862** RAD-CG concrete annotations were obtained and considering multiple annotations for these markers, a total of **3708** annotations are available for them. Among them are **296 CG** with concrete “stress” functions. Many novel CG have been detected which encode concrete genes, but their function is still unknown.

Table 2.2.1i c**o**ntains the first analyses with respect to Specific Expression of RAD markers under Stress Conditions (partial view). A total of **1201/2431 CG** are specifically expressed under drought conditions in susceptible Soprano/ tolerant Kondor, resepctively, **1208/1076** under cold conditions in Soprano/Kondor and **170/972** under heat conditions, using a cut-off value of 10 reads in the stressed genotypes (0 reads in the control). As can be seen from these numbers, apparently more genes are involved in response reactions of tolerant cultivars than in susceptible (as also previously observed in biotic stresses)

In the future also concrete CG alleles which are expressed specifically under stress conditions will be analyzed,
In any case there are sufficient RAD-CG to analyse additional candidate genes for stress tolerance within Task 2.3 in the future.

The EXCEL file ("**"...\PATVIEW\DB\PATLib.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>diff cDNA**.

<u>Period: Month 17 to 24</u>

A second pool of amplicons from the original mRNA populations was produced as described above and was sent for sequencing, in order to increase the total number of sequences for candidate gene detection.

They were processed with in house developed Software as described above.
Sequences of the 2nd run were initially processed by in-house developed Software (**Radlib**) mentioned above, using the options: “Process FastQ”, “Read Through Selector” and “RF Selector” leading to a Y.fasta file of read through fasta sequences and barcodes for the samples. A total of **4,143,191additional sequences** were obtained from the 2nd pool. The Y.fasta files of both runs were combined sequentially and further processed in **Radlib** as described above and in the Software Manual.

After correcting for sequences larger 70 bp a far increased number to **8,476,218 million** useful “read through” sequences were obtained from the digested cDNA libraries of both runs and all the analyses had to be repeated and the corresponding tables described above had to be updated. Using a minimum Threshold of 80 reads per RAD, a total of **12,345 RAD-CG** could be extracted.

Fig.2.2.2.1a (Annex Period Month 17 to 24) shows the characteristics of the experiment including the obtained sequence numbers in each library.

Fig.2.2.2.1b (Annex Period Month 7 to 16) **shows the Reference Sequences of RAD markers in PATLIB (BIOEDIT;** partial view)**.** The sequence lengths range from 55 to 300 bp.

Fig.2.2.2.1c (Annex Period Month 7 to 16) shows the Distribution of Reads of RAD markers over GT. A large degree of variation can be observed.

Fig.2.2.2.1d presents the obtained polymorphisms of the RAD-CG markers (partial view).

For **2896** of the **12345 RAD-CG no** consistent patterns were obtained. A total of **3095 RAD** were monomorphic. The other **6354 RAD-CG** revealed between 2 and 5 patterns in the eight samples which were defined by **1 to 30 SNPs** and in **29** cases more, even up to **62** SNP for quite different alleles (multi-locus CG).

Fig.2.2.2.1e Contains the concrete Allele Sequences, SNP and Pattern Definitions of all RAD markers from PATLib (partial view).

Fig.2.2.2.1f (Annex Period Month 7 to 16) contains the CG Alleles, patterns and SNP definitions in the eight POTATO samples (partial view).

Annotations for the RAD-CG were produced for the RF markers using Mercator Software.
Fig. 2.2.2.1g (Annex Period Month 7 to 16) shows the functional Distribution of the Annotations for RAD-CG marker in PATLib. As can be seen in this picture almost **63%** of the RF could be assigned to specific functions (**37.1% unassigned**), targeting a broad range of functions. Of particular interest are the **4.32%** of RAD markers which are assigned to stress functions.

Fig.2.2.2.1h presents a partial view of the concrete Annotations obtained for the over 12000 RF-CG markers. For **6385** RAD-CG concrete annotations were obtained and considering multiple annotations for these markers, a total of **8566** annotations are available for them. Among them are **116 CG with concrete “stress” functions** and **502 annotations consider “stress”**, Many novel CG have been detected which encode concrete genes, but their function is still unknown.

Fig.2.2.2.1i contains the first analyses with respect to Specific Expression of RAD markers under Stress Conditions (partial view) in the two cultivars and **Table 2.2.2.1j** summarizes the No of Specifically Expressed RAD markers in each library.
In **Soprano** a total of **1814 CG** are specifically expressed under drought conditions, **1709** under cold conditions and **395** under heat conditions, using a cut-off value of 10 reads in the stressed genotypes (0 reads in the control).These numbers are increased for the tolerant cultivar **Kondor**. Many of these genes are expressed under more than one stress condition. They are indicated in the figure with red font and rosé background, while the other stress specific RAD CG are Bold with white background.

In the future also concrete CG alleles which are expressed specifically under stress conditions will be analyzed,
In any case there are sufficient RAD-CG to analyze additional candidate genes for stress tolerance within Task 2.3 (see below).

The **updated** EXCEL file ("**"...\PATVIEW\DB\PATLib.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>diff cDNA**.

Period: Month 25 to 30

While in previous analyses only presence or absence of any CG allele or allele combination under stressed and unstressed conditions was considered, in this period we have analyzed concrete CG alleles or allele combinations which are expressed specifically under stress conditions. That means that the CG is expressed in both situations (stressed and unstressed), but different alleles or allele combinations appear under each condition. Similar results and values can be found for Soprano under cold conditions, while the specific Allele/AC response under heat conditions is lower

The results are shown in Table 3.2.1. A total of 3013 CG show differential expression under drought conditions in the susceptible cultivar Soprano. In 651 cases different alleles appear under drought conditions and unstressed control, while in 2363 cases different allele combinations appear at least in one condition.

For the supposed tolerant variety Kondor, the specific response to stress conditions is much lower compared to Soprano.

A total of 659 CG are differentially expressed under drought conditions in Kondor. Only in 68 cases different alleles appear under drought conditions and unstressed control, while in 591 cases different allele combinations appear at least in one condition.

These numbers are somewhat higher, but similar in their proportions under cols and heat conditions (Table 3.2.1).

The user may check which annotated genes are particularly affected. The results are integrated in the **updated** EXCEL file ("**"...\PATVIEW\DB\PATLib.xlsx"**") under worksheet “SpecAL” which can be accessed in: **PATVIEW>Databases>diff cDNA**.

Period: Month 31 to 39
This Task was completed in Month 30, no further activities.

A 2.2 Analysis of known candidate genes for biotic and abiotic stresses.

Period: First 6 Months

NEIKER has performed in silico mining of sequence databases and publications in order to detect published candidate genes for stress tolerance in potato but also in other species. In this latter case the potato homologs were identified through BLAST searches against the Potato whole genome sequence.

So far a total of 57 candidate genes have been identified. For many of them also the corresponding primers were identified in exon (=coding) regions, since we work later with DNA. All information has been compiled into a CG Database. A partial view is shown in Fig 2.2 of the Annex. This database is also freely available at the project WEB page and forms part of the Knowledge DB on stress resistance in Potato (see below).

Beside the name and primers sequences to obtain amplicons, in many cases Accession numbers, gene sequences and amplicon sequences are indicated in this database.

Period: Month 7 to 16

The CG Stress database has been further expanded with new, known CG for abiotic stresses and resistances to *Phytophthora infestans* and contains now over 120 putative CG for stress and disease resistance (see Table 2.2.2a).

Primers were designed for these CG and are currently tested for functionality

Period: Month 17 to 24

The **CG Stress database** in **PATVIEW** has been further refined and primers were evaluated. The database has been updated with additional potential Resistance genes for abiotic stresses, mainly derived from NCBI resources for Potato and considering Late Blight resistance. In addition, potential resistance genes from the publication: Mosquera et al. 2016. Targeted and Untargeted Approaches Unravel Novel Candidate Genes and Diagnostic SNPs for Quantitative Resistance of the Potato (Solanum tuberosum L.) to *Phytophthora infestans* Causing the Late Blight Disease. Plos One, https://doi.org/10.1371/journal.pone.0156254 were derived.

The database contains now **126** potential Resistance genes. **Primers** were designed for all CG and were tested for **functionality. Fig.2.2.2.2a** presents a partial view of the updated CG Database for abiotic Stresses and ***P. infestans*** resistance in Potato.

Period: Month 25 to 30

The **CG Stress database** has been further enhanced by integrating **47 additional CG** which were evaluated within project PAS4 (see below). These were derived from Specific Expression of RAD markers under Stress Conditions in Task 2.1. The database contains now a total of **216** CG for productivity, quality and resistances.

Primers were designed for all CG and were tested for **functionality. Fig.3.2.2.2a** presents a partial view of the updated CG Database in Potato.

In addition, based on their annotations, some additional RAD-CG markers have been identified which are involved in disease resistance and stress response (see below). They represent good candidates for future applications.

Period: Month 31 to 39

The **CG Stress database** has been further enhanced by integrating **100 additional CG** which were evaluated within project PATF (see below). These were derived from: “Isolated

polynucleotides and polypeptides for increasing plant yield, biomass, growth rate, vigor, starch content, abiotic stress tolerance of plants and nitrogen use efficiency" from different patents by Granevitze, Z. and Karchi,H.
The database contains now a total of **316** CG for productivity, quality and resistances.
Primers were designed for all CG and were tested for **functionality. Fig.4.2.2.2a** presents a partial view of the updated CG Database in Potato.

A 2.3: Analyses of CG by Amplicon Sequencing (CG driven approach)

Period: Month 7 to 16

The CGs which had been detected in Task 2.2 were used for Amplicon sequencing and sequence analyses using the CG driven approach.

The ILLUMINA MySeq Platform was used for sequencing. To our first experiences this platform produces much more sequences of better quality, but the amplicon preparation has to be performed with more accuracy.

A first batch of **45 CG** was processed as project: **PAB1**.
The scheme of the procedure is similar to that one of library construction, shown in **Fig. 2.1 (Period 1)**. However, instead of Restriction site adapters, for the first round of PCR fusion primers, consisting of one part which is complementary to the distal sequences of the expected amplicon of the specific candidate genes and another common part, which is complementary to the index adapters of the Illumina sequencing primers.
In a second round of PCR each sample is re-amplified with the barcoded Index adapters of Illumina. Multiplexing is possible for around each eight CG.
Primers and indexes of ILLUMNA MySeq are shown in **Fig. 2.2.3a**.
The PCR amplicons of each genotype wee purified and mixed in approximately equal amounts. After quality control they were for sequencing (2x300 bp) and the obtained sequences were initially processed as described above in Task 2.1.

They were further processed with *in-house* developed Software (ASPAM) as fastaY files as usual.
This Software is specifically designed to (i) analyze the output of CG Amplicon Sequencing and to extract potential alleles in a set of genotypes and (ii) links allele patterns to their phenotypic expression and analyze their effects through Association Mapping.
The Sample Data consist of: (i) Sequence Data: The sample data consist of large amount of sequence reads from a pool of PCR products of partial DNAs from different candidate genes (CG) in a set of genotypes. These amplicons have forward and reverse adaptors which have genotype specific (barcodes) and (ii) Phenotypic Data for different traits which are available for each Genotype of the set
(Basically the Software performs the following tasks: Extraction of complete sequences and separation for each CG and Genotype; not necessary with Illumina sequences). Determination for each CG the alleles which exist in the set of genotypes and analyses the obtained SNPs and allele patterns. Determination of the allele composition in each sample (genotype). Analysis of effects of alleles, SNPs, allele combination by associating them with the phenotypic performance of the set of genotypes through Association Mapping by using SAS procedures Proc GLM and Proc Stepwise for multiple regression analysis.
Details of the whole process with many alternative options are given in the ASPAM manual.

The Characteristics of the Sequence processing and Association Mapping experiment **PAB1** are shown in **Table 2.2.3b**.
The total Nº of clean reads (= “Read through” sequences) in PAB1 for the **45 CG** was **2,967,287** after removing duplicates (**65,939 per CG**). As usual, the sequence numbers per CG varied considerably as well as those among genotypes.
Two CG did not produce amplicons and 4 CG were found to be monomorphic. The other 39 CG revealed up to **17 patterns** (“alleles”), in one case up to 25 patterns.
SNP numbers varied from **1 to 22** within alleles, beside the one with many different alleles which had **37** SNPs (**Table 2.2.3b**).
Table 2.2.3c contains the list of amplicons as displayed in the Annex (Period month 7 10 16) in BIOEDIT (partial view)**.**

Table 2.2.3d contains the concrete Allele Sequences, SNP and Pattern Definitions in Project PAB1 obtained for each CG (partial view).

Table 2.2.3e shows the Allele composition of all genotypes for all CG (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PAB1-AM.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAB1-AM1 DB**.

A second batch with over 50 new CG for abiotic and disease resistance mentioned in Task 1.2 is in preparation. Illumina MySeq sequencing will be used.

Period: Month 17 to 24

A)

For a total of **29 new CG for Phytophthora Resistance (See A.2)** amplicons were produced in the samples as described above and send for ILLUMINA MySeq sequencing. The obtained sequence reads were added to **PAB1** and the combined sequences were analysed as **new project PAS1. In total now 74 CG were analysed in PAS1**

The total number of sequences in PAS1 were **7,132,310**. A total of 4,260,547 sequences were larger tan 120 bp and 3,344,159 sequences were “Read through” sequences with MIDS at both sides. The Characteristics of the Sequence processing and Association Mapping experiment **PAS1** are shown in **Table 2.2.2.3b**.
As usual, the sequence numbers per CG varied considerably as well as those among genotypes.
Two of the 74 CG did not produce amplicons and **10 CG** were found to be monomorphic. The other **62 CG** revealed up to **17 patterns** (“alleles”).
SNP numbers varied from **1 to 24** within alleles, beside one with quite different alleles which had 43 SNPs (**Table 2.2.2.3b**).
Fig. 2.2.2.3c contains a partial view of the list of expected amplicons in PAS1.

Fig. 2.2.2.3d contains the concrete Allele Sequences, SNP and Pattern Definitions in Project PAS1 obtained for each CG (partial view).

Fig. 2.2.2.3e shows the Allele composition of all genotypes for all CG (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PAS1car.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAS1-AM**.

B)

A second batch of **43 new CG** for resistances mainly from NCBI resources (See **A.2**) was processed as project: **PAS2**. All Potato samples were prepared for Amplicon sequencing using the ILLUMINA MySeq Platform as described above. The received sequences were processed using ASPAM Software as described above.

The Characteristics of the Sequence processing and Association Mapping experiment **PAS2** are shown in **Table 2.2.2.3f**.
The total Nº of clean reads (= “Read through” sequences) in PAS2 for the **43 CG** was **6,116,510** after removing duplicates (**142,244 per CG**). As usual, the sequence numbers per CG varied considerably as well as those among genotypes.
Five CG did not produce amplicons and **3 CG** were found to be monomorphic. The other **35 CG** revealed up to **21 patterns** (“alleles”), in one case up to 25 patterns.
SNP numbers varied from **1 to 24** within alleles, beside one with quite different alleles which had 40 SNPs (**Table 2.2.2.3f**).
Fig. 2.2.2.3g contains a partial view of the list of amplicons of PAS2.

Fig. 2.2.2.3h contains the concrete Allele Sequences, SNP and Pattern Definitions in Project PAS2 obtained for each CG (partial view).

Fig. 2.2.2.3i shows the Allele composition of all genotypes for all CG in PAS2 (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PAS2car.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAS2-AM**.

C)
From the first PATLib library promising resistance CG were extracted and converted to RAD-CG markers by blasting against the potato chromosomes. In part they were further extended with adjacent sequences of the genome location. A total of **42 RAD CG** were processed in this way and amplicons were generated as described above and send for sequencing on a Illumina MySeq Platform. The obtained sequences were processed as project **PAS3.**

The Characteristics of the Sequence processing and Association Mapping experiment **PAS3** are shown in **Table 2.2.2.3j.**
The total Nº of clean reads (= “Read through” sequences) in **PAS3** for the **42 CG** was **3,098,021** after removing duplicates (**73,762 per CG**). As usual, the sequence numbers per CG varied considerably as well as those among genotypes.
Four CG did not produce amplicons and **2 CG** were found to be monomorphic. The other **36 CG** revealed up to **14 patterns** (“alleles”).
SNP numbers varied from **1 to 25** within alleles, beside one with quite different alleles which had 35 SNPs (**Table 2.2.2.3j**).
Fig. 2.2.2.3k contains a partial view of the list of amplicons of **PAS3**.

Fig. 2.2.2.3l contains the concrete Allele Sequences, SNP and Pattern Definitions in Project **PAS3** obtained for each CG (partial view).

Fig. 2.2.2.3m shows the Allele composition of all genotypes for all CG in **PAS3** (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PAS3car.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAS3-AM**.

Period: Month 25 to 30

A fourth batch of **47 new CG** for resistances (See **A.2**) was processed as project: **PAS4**.

All Potato samples were prepared for Amplicon sequencing using the ILLUMINA MySeq Platform as described above. The received sequences were processed using ASPAM Software as described above.

The Characteristics of the Sequence processing and Association Mapping experiment **PAS4** are shown in **Table 3.2.2.3b**.
The total read number was **4,082,456** sequences >120 bp in the experiment.
The total Nº of clean reads (= "Read through" sequences) in PAT2 for the **47 CG** was **2,971,417** after removing duplicates (**63,222 per CG**). As usual, the sequence numbers per CG varied considerably as well as those among genotypes. **25 CG** did not produce sufficient amplicons to derive patterns. The other **22** CG revealed up to **14 patterns** ("alleles").
SNP numbers varied from **1 to 25** within alleles and 48 in one case for quite different alleles (T**able 3.2.2.3b**). **Fig. 3.2.2.3c contains a partial view of the list of amplicons in PAS4.**

Fig. 3.2.2.3d contains the concrete Allele Sequences, SNP and Pattern Definitions in Project **PAS4** obtained for each CG (partial view).
Fig. 3.2.2.3e shows the Allele composition of all genotypes for all CG in PAS4 (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PAS4car.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results…>PAS4-AM**.

Period: Month 31 to 39

In this period a fifth batch of **101 new CG** derived from different patents for increasing plant yield, biomass, growth rate, vigor, oil content, abiotic stress tolerance of plants and nitrogen use efficiency (See **A.2**) was processed as project: **PATF**.

All Potato samples were prepared for Amplicon sequencing using this time the ION Torrent Sequencing Platform with the new chip I-530. Using this chip up to 25 million reads of 400 bp can be obtained in a cost efficient way. The scheme for producing the barcoded amplicons is identical to that one of ILUMINA MySeq, but different, platform specific universal and barcode primers (MIDs) are used. Moreover, no separate forward and reverse reads for merging, but one unidirectional read is provided for each sequence. The received sequences were processed as usual using ASPAM Software.

The Characteristics of the Sequence processing and Association Mapping experiment **PATF** are shown in **Table 4.2.2.3b**.
The total read number was **15,113,305** sequences >120 bp in the experiment.
The total Nº of clean reads (= "Read through" sequences) in PATF for the **101 CG** was **11,781,007.** A total of **8,676,079** sequences could be assigned to the **101 CG** after removing duplicates (**85,902 per CG**). As usual, the sequence numbers per CG varied considerably as well as those among genotypes. **11 CG** did not produce sufficient amplicons to derive patterns, **4** CG were monomorphic. The other **86** CG revealed up to **18 patterns** ("alleles").
SNP numbers varied from **1 to 46** within alleles of sometimes quite different alleles (T**able 4.2.2.3b**). **Fig. 4.2.2.3c contains a partial view of the list of amplicons in PATF.**

Fig. 4.2.2.3d contains the concrete Allele Sequences, SNP and Pattern Definitions in Project **PAS4** obtained for each CG (partial view).
Fig. 4.2.2.3e shows the Allele composition of all genotypes for all CG in PAS4 (partial view).

The EXCEL file ("**"...\PATVIEW\DB\PATFcar.xlsx"**") with all results has been integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results…>PATF-AM**.

A 2.4: RAD sequencing for CG detection and analyses (random approach).

Period: Month 7 to 16

RAD sequencing will allow to target a large number of genes, since partial cDNAs representing coding regions are analysed. Among them are an elevated number of potential CGs for the analysed traits, based on their biological meaning.
The process is similar to that one of Task 2.1, but extended to a large number of genotypes. The Methodology has been established successfully in the frame of other R&D projects at NEIKER. Also the necessary software (RADSAT has been developed updated for this purpose and is available for the analyses.

The necessary materials (either RNA or samples in RNA later of each genotype) are in preparation by the partners. NEIKER is waiting to receive them in order to start immediately with the processing.

Period: Month 17 to 24

Leaf samples in RNA latter were received from the partners in Ecuador and Peru MARI from their corresponding accessions. RNA was extracted from each sample and was processed for ILLUMINA MySeq sequencing in the same way as described in A2.1 (Library construction). Each sample gets a combination of specific I5 and I7 indexes.
In this reporting period one pool has been prepared and send for sequencing very recently (July 10,218). Results are pending. For the near future the preparation and analysis of two additional pools is foreseen in order to increase the total RAD CG number.

Period: Month 25 to 30

In this reporting period RAD sequencing has been completed. Two additional pools were prepared as described above in order to increase the total RAD CG number. The sequences of all three pools were combined and analysed jointly.

From the three combined pools initially **over 18 million sequences** were obtained. After cleaning a total of **7,907,301 “read through” sequences** assigned to RAD-CG remained which were processed by in-house developed Software (Radlib/Radsat) as described before. A total of **675 RAD-CG** markers could be extracted and were processed as project **RPAT**.

Fig. 3.2.4a presents a partial view of the Sequences of the 675 RF-CG markers in Fasta Format
The sequence lengths range from 100 to over 300 bp in the RAD-CG.

Fig. 3.2.4b presents a partial view of the obtained polymorphisms in the 675 RAD-CG markers.
A total of 31 RAD-RF were monomorphic and 20 RAD-RF did not produce consistent reads. The other **623 RF-CG** revealed between 2 and 38 patterns in the population which were defined by 1 to 54 SNPs.

Annotations for the RF-CG were produced for the RF markers of RPATusing Mercator Software.
Fig. 3.2.4c shows the functional Distribution of the Annotations for RF-CG markers. As can be seen in this picture almost 58% of the RF could be assigned to specific functions (42.2% unassigned), targeting a broad range of functions.
Fig. 3.2.4d presents a partial view of the concrete Annotations obtained for the 675 RF-CG markers in RPAT. A total of **504 annotations** were obtained for the 675 markers (75%).

RF-CG markers were also blasted against the POTATO chromosome sequences and the annotated genes on these chromosomes.
Fig. 3.2.4e presents a partial view of the Map locations for the 675 RF-CG markers on the chromosomes. A total of **362 RF could be mapped** to the Potato genome covering all chromosomes (54%).
Fig. 3.2.4f presents a partial view of RAD-CG markers which could be mapped to annotated Genes on the chromosomes. A total of **336 previously annotated Genes** on these chromosomes were targeted by RF markers (50%).
The EXCEL file ("**E:\VIEWER\PATVIEW\DB\RadPAT.xlsx**") with all UPDATED results can be accessed in: **PATVIEW>DATABASES>AM-Results>RAD DB**.

Period: Month 31 to 39
This Task was completed in Month 30, no further activities.

Specific Objective 3: Association Mapping and Model Development

A 3.1- Association Mapping

Period: Month 7 to 16

In this task the molecular data are linked to the phenotypic data and for each CG the effects of alleles and allele combination on trait expression is determined.
The molecular data have been prepared in Task 2.3 and are ready for use. Also the necessary software (ASPAM) has been updated for this purpose and is available for the analyses.

The phenotypic data are in preparation by the partners. NEIKER is waiting to receive them in order to start immediately with the processing.

Period: Month 17 to 24
Phenotypic Records for different traits were received from the partners from Ecuador and Peru from their corresponding accessions and field trials.

The trait data received so far consider: tuber number (**NT**), average tuber weight (**ATW**), tuber yield (**Y**), cold damage (tolerance, **DH**) and Phytophthora tolerance data expressed as AUDC values (**AUD**).

For processing yield and yield component data from different experiments and different plant materials, these were converted into relative values with respect to the population mean of each trial (=100%)!

Fig. 2.3.1a presents a partial view of the combined phenotypic trait records in the potato accessions.

For all traits large variation can be observed in the collection. Also the detected correlations are as expected between competing organs or based on physiological relationships.

Association mapping to determine Allele and AC effects on the provided Traits were performed in projects PAS1, PAS2 and PAS3. Numerous effects were detected.
Fig. 2.3.1b presents a partial view of Pattern (=allele) effects and SNP effects in PAS1 for the available Traits.
Fig. 2.3.1c presents a partial view of the detected Allele combination effects and Homo- vs Heterozygous effects in PAS1 for the available traits.
Fig. 2.3.1d summarizes all results from Association Mapping in PAS1.

For **24 CG** out of 74 significant allele or SNP effects were observed in PAS1, 13 in pattern analyses involving 1 to 2 traits and **24** in SNP analyses involving also 1 to 3 traits.
In addition, in PAS1 **16** significant allele combination effects were observed involving up to 2 traits. In Ho/He Analyses **5** CG showed significant differences involving 1 to 2 traits.

The association mapping results were much better In PAS2:
Fig. 2.3.1e presents a partial view of Pattern (=allele) effects and SNP effects in **PAS2** for the available Traits.
Fig. 2.3.1f presents a partial view of the detected Allele combination effects and Homo- vs Heterozygous effects in **PAS2** for the available traits.
Fig. 2.3.1g summarizes all results from Association Mapping in **PAS2**.
For **26 CG** out of 43 significant allele or SNP effects were observed in PAS2, 24 in pattern analyses involving 1 to 4 traits and **26** in SNP analyses involving also 1 to all 5 traits.
In addition, in PAS2 **21** significant allele combination effects were observed involving up to 4 traits. In Ho/He Analyses **14** CG showed significant differences involving 1 to 2 traits.

The analogous Association Mapping results are also available for the CG from PAS3.
Fig. 2.3.1h presents a partial view of Pattern (=allele) effects and SNP effects in **PAS3** for the available Traits.
Fig. 2.3.1i presents a partial view of the detected Allele combination effects and Homo- vs Heterozygous effects in **PAS3** for the available traits.
Fig. 2.3.1j summarizes all results from Association Mapping in **PAS3**.
For **30 CG** out of 42 significant allele or SNP effects were observed in PAS3, 25 in pattern analyses involving 1 to 3 traits and **30** in SNP analyses involving also 1 to 4 traits.
In addition, in PAS3 **20** significant allele combination effects were observed involving up to 4 traits. In Ho/He Analyses **10** CG showed significant differences involving 1 to **3** traits.

The EXCEL file ("**"...\PATVIEW\DB\PAS1car.xlsx"**") with all **PAS1 results** has been also integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAS1-AM.**
In an analogous way the results from **PAS2** can be accessed at: **PATVIEW>Databases>AM Results>PAS2-AM** and those from **PAS3** at **PATVIEW>Databases>AM Results>PAS3-AM**

Period: Month 25 to 30

Association mapping in PAS4

Association mapping to determine Allele and AC effects on the provided Traits were performed in project PAS4 with the initial data**. Numerous effects were detected.**

Fig. 3.3.1a presents a partial view of Pattern (=allele) effects and SNP effects in PAS4 for the available Traits. A total of **10** CG allele effects and **48** SNP effects were detected.

Fig. 3.3.1b presents a partial view of the detected Allele combination effects and Homo- vs Heterozygous effects in PAS4 for the available traits. A total of **19** CG allele combination effects and **12** Ho vs Het effects were detected.

Fig. 3.3.1c summarizes all results from Association Mapping in PAS4.
For 11 CG out of 22 polymorphic significant allele or SNP effects were observed in PAS4, 9 in pattern analyses involving 1 or 2 traits and **12** in SNP analyses involving up to 4 traits.
In addition, in **PAS4 for 12** CG significant allele combination effects were observed involving up to 4 different traits. In Ho/He Analyses **9** CG showed significant differences involving 1 to 3 traits.

The EXCEL file (""**...\PATVIEW\DB\PAS4car.xlsx**"") with all **PAS4 results** has been also integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>PAS4 DB**.

Association mapping in PATR

In an analogous way also Association mapping of the RAD-CG markers of **Project PATR** with the available trait Data were performed, and numerous associations were obtained as usual.

Fig. 3.3.1d presents a partial view of Pattern (=allele) effects and SNP effects in PATR for the available Traits. A total of **406** CG allele effects and **1346** SNP effects were detected

Fig. 3.3.1e presents a partial view of the detected Allele combination effects and Homo- vs Heterozygous effects in PATR for the available traits. A total of **210** CG allele combination effects and 116 Ho vs Het effects were detected for the 8 traits.

Fig. 3.3.1f summarizes all results from Association Mapping in PATR.
For 354 CG significant allele or SNP effects were observed in PATR, **239** in pattern analyses involving up to 5 traits and **349** in SNP analyses involving 1 to 4 traits.
In addition, in **PATR for 172 CG** significant allele combination effects were observed involving up to 3 different traits. In Ho/He Analyses **92** CG showed significant differences involving also 1 to 3 traits.

The EXCEL file (""**...\PATVIEW\DB\RADPAT.xlsx**"") with all **PATR results** has been also integrated in **PATVIEW** and can be accessed in: **PATVIEW>Databases>AM Results>RAD DB.**

Period: Month 31 to 39

New and updated phenotypic data records were received from the partners for many traits (see below).
The PATF project was directly combined with the previous PATX project and Association mapping was performed within T3.4 Model building in order to not duplicate the results (see below).

A 3.2 Design of allele specific primers (ASP) for important CG alleles

Period: Month 17 to 24

We have initiated allele specific primer (**ASP**) design using the following strategy:
We have searched for CG for important traits and/or which explain a large amount of the total variance. These data are available in the Excel worksheets with association mapping results integrated in PATVIEW (AM Results for Allele Models).
For this purpose the markers with significant Allele effects were ordered by trait and significance values. We intended to design ASP for CG alleles with highest significances (generally <.0001, or other very small values).

In this way an initial list of **43** promising CG alleles for the 5 traits were established. They are shown together wItrh their characteristics in **Table 2.3.2a** (left side).

Then the selected CG alleles were traced back in the original projects (PAS1, PAS2) where they descended from.
The SNP and Pattern definitions of all alleles belonging to the CG of the specific allele of interest were inspected carefully in order to detect SNP level values which were unique for this allele among all other alleles.

This was the case for **20** alleles out of the 43 selected. For each CG allele the descend Project, the original CG number, the specific SNP number, the level value and the alternative allele in IUPAC code are also shown in Table 2.3.2a on the right side.
In the future it should be also possible to design ASP for additional CG alleles, but using a combination of different "excluding" primers. The combination of the individual amplification events would identify the specific allele of interest.

For designing the ASP the Software BatchPrimer 3 (https://probes.pw.usda.gov/batchprimer/, Albany Server) was used. This Software requires the generation of a fasta file in a specific fomat. It is shown in Fig. 2.3.2b and was prepared manually, based on the information of the "CG Allele SNP file" of the corresponding project.
Batch primer 3 detected for 17 of the 20 selected CG alleles specific primers. They are displayed in Fig.2.3.2c. Details about primers and expected amplicons for **ASP18** are shown in Fig. 2.3.2d.
ASP could be designed for several important markers until now as shown in **Fig. 2.3.2a**. These include for example **ASP42** for A1-EFF(254BP) which increases tuber yield by almost 50%, if present. Notice that several different traits may be influenced by the same allele as it is here the case for tuber number.
Another marker is **ASP18** for A10-ABCG1-40(298BP, with a tremendous effect on the AUDPC values. If present, it reduces it to almost 50%.

The complete results with all details can be accessed via PATVIEW>Databases>Allele Specific Primers>ASP1. This link will open the ASP subdirectory with all files. **Fig. 2.3.2e** shows the contents of the ASP sub directory.

Period: Month 25 to 30

ASP primer design has been continued in this period. This time we have targeted the RAD markers of RPAT. This time the strategy was slightly different.
We have searched for CG for the available traits which have highly significant allele effects and identified the "best" alleles, which are significantly better than other alleles. Notice, that usual these are alleles with highest values, except for "DH" and "AUD" where smallest values are desired. These data are available in the Excel worksheets with association mapping results integrated in PATVIEW (AM Results for Allele Models).
Then the selected CG alleles were traced back in project **RPAT** where they descended from. The SNP and Pattern definitions of all alleles belonging to the CG of the specific allele of interest were inspected carefully in order to detect SNP level values which were unique for this allele among all other alleles.
If this was the case, the Software BatchPrimer 3 (https://probes.pw.usda.gov/batchprimer/, Albany Server) was used for designing the ASP. This Software requires the generation of a fasta file in a specific format. and was prepared manually, based on the information of the "RAD-CG Allele SNP file" of the corresponding project.
If ASP were obtained for the specific CG allele, the Candidate CG and alleles was added to **Table 3.3.2a** which contains a list of RAD-CG, alleles and characteristics for which ASP could be designed.
In this way a list of 19 promising RAD-CG alleles for the different traits were established (T**able 3.3.2a; ASP-CG.xlsx**). Notice, that sometimes the same CG has effects on different traits. These cases are in included in Table 3.3.2a, without yellow shading.
The Fasta file for Batchprimer 3 is shown in **Fig. 3.3.2b (ASP2.txt)** for all RAD-CG alleles for which ASP could be designed. The name of each sequence identifies directly the CG, the particular allele and the targeted trait. Notice, that also negative selection may be possible.

For example **“>P92-no2AUD**” means that allele 2 of P92, should not be present, since it increases the AUD values considerably.

The 19 detected CG alleles specific primers and characteristics are displayed in Fig.3.3.2c (file:///E:/VIEWER/PATVIEW/ASP2/Primer_table.html).
Clicking on **a SEQ ID** will display details about primers and expected amplicons of the **ASP** (**Fig. 3.3.2d**).

Additional ASP could be designed for several important markers as shown in **Fig. 3.3.2a**. These include for example the four **ASP** for “DH· (Cold tolerance) which amplify only in the highly tolerant accessions (PE184 and PE185). Other ASP such as P38 allele 1 increase tuber yield and tuber weight by over 100%,
Another marker such as P616 allele 2 has a tremendous effect on the AUDPC values. If present, it reduces it to almost 150%.

Together with the 18ASP from the previous reporting period, a total of 35 ASP are now available, more than required based on the Work plan (30 ASP)

The complete results with all details can be accessed via **PATVIEW>Databases>Allele Specific Primers>ASP2**. This link will open the ASP subdirectory with all files. **Fig. 3.3.2e** shows the contents of the ASP2 sub directory.

Period: Month 31 to 39
Within Task A3.4 (see below) SNP were extracted from the PASi and PATF projects and from the RAD markers and so called SNP projects PATXS and PATRS, respectively, were established. These were used to design a final set of ASP markers for the traits AUD (*Phytophthora* resistance) and DF8 (Frost tolerance at -8ºC) which are the most important traits for the partners.
SNP projects present an alternative method to circumvent the problem of dosage effects and multi locus CG, since SNP levels are often shared by more than one allele and may be locus specific. Also the design of ASP markers for different SNP levels may be more convenient than ASP primers for presence or absence states of specific alleles.

SNP projects derived from CG and RAD markers were processed in the same way as CG Allele projects. For AUF and DF8 the effects of SNP levels (reference and alternative nucleotides) and SNP level combinations including the heterozygous states were determined.
We have selected those SNP with largest effects explaining large amount of the total variance (>20%) for the mentioned traits and designed allele specific primers as usual.

The selected SNP were traced back in the projects **RPAT** or **PASi**, respectively where they originally descended from.
The SNP and Pattern definitions of all alleles belonging to the specific SNP of interest were inspected carefully in order to detect SNP level values which were unique for this SNP among all other SNPs.
If this was the case, the Software BatchPrimer3 (https://probes.pw.usda.gov/batchprimer/, Albany Server) was used for designing the ASP for the selected SNPs. This Software requires the generation of a Fasta file in a specific format. and was prepared manually (**Table 4.3.2a)**, based on the information of a pattern where the specific SNP was present from the corresponding project.
If ASP were obtained for the specific SNP was added to **Table 4.3.2b** which contains a list of SNP and characteristics for which ASP could be designed.

SNP Primers for **AUD** from CG have names ASP3-x, and from RAD : ASP3R-x, SNP Primers for **DF8** from CG have names ASF3-x, and from RAD : ASF3R-x
Table 4.3.2c shows a partial view of the output from BatchPrimer3. For a total of **20** SNP ASP primer pairs could be found, sometimes more than one, leading to a total of **80** primers. Clicking on a Seq ID, shows the details of the primers for the selected ASP (**Table 4.3.2d**).

A final list of 19 promising ASP SNP primers for the two traits were established (T**able 3.3.2a; ASP3.xlsx**); 11 for AUD and 8 for DF; 13 descended from PATXS and 6 from PATRS.
The results can be accessed in **PATVIEW> Allele specific Primers > ASP3.**

A 3.3 Model Development

A 3.4 Model Validation and Refinement

Period: Month 17 to 24

Model building has been initiated with the POTATO CG data. Using ASPAM, the PAS1 project was combined with PAS2 as a so called “**external Project”: PAX**.

Since we have in the genotypes a mix of diploid, triploid and tetraploid genotypes and also the dosage (simplex, duplex, triplex, cuadriplex) of a particular allele is difficult to determine, all alleles of each CG were converted into “Allele Markers” with the levels present (=1) versus absent (=2) and the data were analyzed as usual for association mapping and Model building in ASPAM.

Notice that with this type of presence-absence Data only ALLELE models can be computed. Also the “CG characteristics” have to be generated artificially, since they are needed for the functioning of the Software. The particular CG alleles have to be traced back in the original databases of PAS1, PAS2 and PAS3 respectively.

Fig. 2.3.3a contains a partial view of the CG Characteristics of Project**: PAX**. A total of **242** Allele markers were available in this way..

Fig. 2.3.3b shows partially the molecular patterns of CG markers in the available samples of PAX.

The Phenotypic DATA available for the PAX project were the same as for the other Projects PAS1 to 3.
Fig. 2.3.3c shows a partial view of the AM Results of CG allele analysis (Allele effects) in PAX Numerous highly significant associations were obtained as usual. A total of **175** significant effects of CG alleles were obtained involving up to all 5 traits. Obviously the results of SNP analyses are the same for this data type (**Fig. 2.3.3d**)..

Fig. 2.3.3e shows the Model analyses results for all traits in PAX. Correlations between predicted and observed performances ranged **between 37 and over 70%**, explaining up to **50%** of the total variance. The correlations for the Trait DH (cold damage) were not significant since the distribution is very skewed (non normal distribution) and only a very few cold tolerant genotypes were detected.
Fig. 2.3.3f presents a partial view of the **breeding values** (BV) of the genotypes in **PATX** if used for breeding

Fig. 2.3.3g presents a partial view of the prediction matrixes of progeny performances for YIELD of all crosses between genotypes derived from AL models. GT are ordered for BV, so that the TOP crosses (highlighted) can be identified easily.
The EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATX.xlsx**") with all mentioned results can be accessed in: **PATVIEW>DATABASES>AM-Results>PATX DB**.

Period: Month 25 to 30

1. MODEL BUIDING in updated PATX

Updated phenotypic Data were received from the partners. For Peru these consisted mainly of cold tolerance data from experiments under controlled conditions (-4º and -8ºC; DF4 and DF8). For Ecuador these included mainly updates in the AUDPC data. In addition some comparative data for the performance under different stress conditions were provided for a reduced number of genotypes. While these data are useful for physiological studies, there are not sufficient accessions for Association mapping and Model building.

The mentioned updated data records were used for Model building. The updated Phenotypic DATA are the same for all Potato Projects. Data records for 7 traits were available. **Fig. 3.3.3a** contains a partial view of the updated phenotypic trait values.

Using ASPAM, the PAS1, PAS2, PAS3 and PAS4 projects were combined as **UPDATED**, **external Project PATX**. CG markers were filtered for Reads in more than 80 genotypes per CG, all **CG** alleles were converted into **Multi Locus Fragments.** In addition, missing values were computed and the data were analyzed as usual for association mapping and Model building.

Fig. 3.3.3b contains a partial view of the CG Characteristics of the updated Project**: PATX**. A total of **747** MLF (=CG allele) markers were available for the analyses.

Fig. 3.3.3c shows partially the molecular patterns of CG allele markers of **PATX** in the Collection,

Fig. 3.3.3d shows a partial view of the AM Results of CG allele analysis (Allele effects) in PATX. A total of 2782 significant associations were obtained for the 7 traits. Since only MLF markers exist the results for AC Models are the same!

Fig. 3.3.3e shows the Model analyses results for all traits. Significant Correlations between predicted and observed performances of **between 52 and 83% were obtained** which explain **between 27 and 69% of the total variance**. Notice that the results are identical for AC models

Fig. 3.3.3f presents a partial view of the breeding values of the genotypes for all traits in **PATX** if used for breeding and **Fig. 3.3.3g** presents a partial view of the prediction matrixes of progeny performances for crosses between PATX genotypes for the selected traits derived from AL models. The user can derive the TOP crosses for each trait by selecting the highest values from this matrix.
Fig. 3.3.3h presents a partial view of the mean performances for all Traits of the genotypes in **PATX** if used for breeding and **Fig. 3.3.3i** presents a partial view of the prediction matrixes of progeny performances for crosses between PATX genotypes for all traits derived from MP values. Notice that the Progeny Performance Prediction values are the same as those for AL models.

The updated EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATX.xlsx**") with all mentioned results can be also accessed in: **PATVIEW>DATABASES>AM-Results>PATX DB**.

2. MODEL BUIDING with Potato RAD markers

Model building was also performed with the Potato RAD markers and the **updated phenotypic Data**. For this purpose RAD-CG markers from project **PATR** were filtered for Reads in more than 80 genotypes per CG, all **CG** alleles were converted into **Multi Locus Fragments.** In addition, missing values were computed and the data were analyzed as usual for association mapping and Model building as project **PatRX.**

Fig. 3.4.3a contains a partial view of the CG Characteristics of Project **PATRX**. A total of **456** MLF (=RAD-CG alleles) markers were available for the analyses and **Fig.3.4.3b** presents a partial view of the RAD-CG allele sequences in **PATRX**.

Fig. 3.4.3c shows partially the molecular patterns of RAD-CG allele markers of **PATRX** in the Collection.

Fig. 3.4.3d shows a partial view of the AM Results of CG allele analysis (Allele effects) in **PATRX**. A total of **2562 significant associations** were obtained for the 7 traits. Since only MLF markers exist the results for AC Models are the same!

Fig. 3.4.3e shows the Model analyses results for all traits in PATRX. Significant Correlations between predicted and observed performances of **between 55 and 78% were obtained** which explain **between 29 and 61% of the total variance**. Notice that the results are identical for AC models

Fig. 3.4.3f presents a partial view of the breeding values of the genotypes for all traits in **PATRX** if used for breeding and **Fig. 3.4.3g** presents a partial view of the prediction matrixes of progeny performances for crosses between PATRX genotypes for the selected traits derived from AL models. The user can derive the TOP crosses for each trait by selecting the highest values from this matrix.
Fig. 3.4.3h presents a partial view of the mean performances for all Traits of the genotypes in **PATRX** if used for breeding. The Progeny Performance Prediction values of AC models are the same as those for AL models.

The updated EXCEL file ("**E:\VIEWER\PATVIEW\DB\RADPAT.xlsx**") contains also the results of Model building in **PATRX** with all mentioned results can be accessed in:
PATVIEW>DATABASES>AM-Results>RAD DB (Worksheets with “PatRX and followings).

Period: Month 31 to 39

1. MODEL BUIDING in updated PATXf

Updated phenotypic Data were received from all partners. For IBT (Peru) these considered updated data for frost tolerance from experiments under controlled conditions (-4º and -8ºC; DF4 and DF8) and field data for tuber number, average tuber weight and tuber yield per plant (Ntpe, ATWpe, Ype). For INIAP (Ecuador) these considered updated field data for tuber number, average tuber weight and tuber yield per plant (NT, ATW, Y), updates in the AUDPC data and as new trait Lesion growth after PI infections (TCL). In addition some comparative data for the performance under drought conditions and control in the greenhouse were provided (Ntis, Ntic, Twis ,Twic, Yis, Yic). **Table 4.3.1a** shows a partial view of the final Phenotypic Data records INIAP and IBT and **Table 4.3.1b** the legend for the trait names for these data. A total of **16 traits** were available from INIAP and IBT.

Partner USFQ performed stress assays under drought, cold and heat conditions using a reduced set of EC accessions. The available trait data consider tuber yield and dry matter in the field (multiplication phase), production traits under stress conditions and control and the production after transplanting the stressed and unstressed materials to the field. In addition some visual damage data and indirect parameters were measured (SPAT, proline contents, Stomata conductance and fluorescence). **Table 4.3.1c** shows a partial view of the final Phenotypic Data records from USFQ and **Table 4.3.1d** the detailed legend for the traits of the USFQ data. **30 traits** were available from USFQ summing up to a total of **46 traits** for the analyses.

NEIKER has used the final data records for all 46 mentioned traits for Association mapping and Model building. Using ASPAM, the PAS1, PAS2, PAS3, PAS4 and PATF projects were combined as **Project PATXf**. CG markers were filtered for Reads in more than 80 genotypes per CG, all **CG** alleles were converted into **Multi Locus Fragments.** In addition, missing values were computed and the data were analyzed as usual for association mapping and Model building.

Fig. 4.3.3b contains a partial view of the CG Characteristics of the updated Project**: PATXf**. A total of **1203** MLF (=CG allele) markers were available for the analyses.

Fig. 4.3.3c shows partially the molecular patterns of CG allele markers of **PATXf** in the Collection,

Fig. 4.3.3d shows a partial view of the AM Results of CG allele analysis (Allele effects) in PATXf. A total of **25359 significant associations** were obtained for the **46** traits. Since only MLF markers exist the results for AC Models are the same! Fig. **4.3.3e** summarizes the AM Results of CG allele analysis for all available traits.

Fig. 4.3.3f shows the Model analyses results for all traits. Significant Correlations between predicted and observed performances are higher in the PATXF model which includes also the markers of PATF and range **between 62 and 85%,** which explain **between 38 and 72% of the total variance**. Notice that the results are identical for AC models, since only presence or absence states of the alleles are considered (= 2 AC classes).

Fig. 4.3.3g presents a partial view of the breeding values of the genotypes for all traits in **PATXf** if used for breeding and **Fig. 4.3.3h** presents a partial view of the prediction matrixes of progeny performances for crosses between PATXF genotypes for the selected traits derived from AL models. The user can derive the TOP crosses for each trait by selecting the highest values from this matrix. Since AL and AC models are identical it makes no sense to compute mean performances and derive the corresponding PPP matrixes.

Instead, we have determined the Most efficient markers by Multiple regression analyses and extracted the top crosses for each of the 46 traits.

Fig. 4.3.3i presents a partial view of a reduced set of most efficient markers derived from AL models as obtained by Multiple Regression explaining a large amount (around 90%) of the total variance.

Fig. 4.3.3j presents a partial view of the top crosses for each trait which were extracted from the PPP matrixes of Allele Models.

The EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATXF.xlsx**") with all mentioned results can be also accessed in: **PATVIEW>DATABASES>AM-Results>PATXF DB** (and replace the previous PATX DB).

2. MODEL BUIDING with Potato RAD markers and final phenotypic Data

Model building was also performed with the Potato RAD markers and the **final phenotypic Data as Project PATRf**. **That means, the molecular data are the same as those of the previous project PATR (456 RAD-CG MLF) and were presented already in the previous report (see**

above). Only the Association Mapping and Model building results have changed and are presented here.

Fig. 4.3.4d shows a partial view of the AM Results of CG allele analysis (Allele effects) in **PATRf.** A total of **10948 significant associations** were obtained for the **46** traits. Since only MLF markers exist, the results for AC Models are the same! Fig. **4.3.3e** summarizes the AM Results of CG allele analysis for all available traits.

Fig. 4.3.4f shows the Model analyses results for all traits. Significant Correlations between predicted and observed performances range **between 44 and 86%** and explain **between 17 and 74% of the total variance**. Notice that the results are identical for AC models, since only presence or absence states of the alleles are considered (= 2 AC classes).

Fig. 4.3.4g presents a partial view of the breeding values of the genotypes for all traits in **PATRF** if used for breeding and **Fig. 4.3.4h** presents a partial view of the prediction matrixes of progeny performances for crosses between PATRF genotypes for the selected traits derived from AL models. The user can derive the TOP crosses for each trait by selecting the highest values from this matrix. Since AL and AC models are identical it makes no sense to compute mean performances and derive the corresponding PPP matrixes.

Instead, we have determined the Most efficient markers by Multiple regression analyses and extracted the top crosses for each of the 46 traits.

Fig. 4.3.4i presents a partial view of a reduced set of most efficient markers derived from AL models as obtained by Multiple Regression explaining a large amount (around 90%) of the total variance.

Fig. 4.3.4j presents a partial view of the top crosses for each trait which were extracted from the PPP matrixes of Allele Models.

The EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATRf.xlsx**") contains also the new results of Model building in **PATRf** with all mentioned results can be accessed in:
PATVIEW>DATABASES>AM-Results>PATRf DB (and replace the previous PATR DB).

3. MODEL BUIDING with SNP projects derived from CG and RAD markers

In this period we have also the SNP extracted from the PASi and PATF projects on one side and from the RAD markers on the other side. Only biallelic and non redundant SNP were extracted. In addition SNP markers were filtered for Reads in more than 80 genotypes per SNP, missing values were computed **and so called** **SNP projects: PATXS and PATRS, respectively, were established.**

SNPs of PATXS and PATRS were used to perform Association mapping and Model building for the traits AUD (*Phytophthora* resistance) and DF8 (Frost tolerance at -8ºC) which are the most important traits for the partners.
SNP projects present an alternative method to circumvent the problem of dosage effects and multi locus CG, since SNP levels are often shared by more than one allele and may be locus specific.

PATXS:

Fig. 4.3.5a contains a partial view of the SNP Characteristics of Project**: PATXS**. A total of **722 nr** SNP markers were available for the analyses.

Fig. 4.3.5b shows partially the molecular patterns of SNP markers of **PATXS** in the Collection,

Fig. 4.3.5c presents a partial view of the Phenotypic Data for **AUD** and **DF8**, which were extracted from the **final phenotypic Data**.

Fig. 4.3.5d shows a partial view of the AM Results of SNP allele analysis (Allele effects) in PATXS. A total of **757 significant associations** were obtained for the **2** traits.

Unlike in PATXF, AL and AC models are different here, since also heterozygous SNP genotypes can exist for each SNP.

Therefore, Fig. 4.3.5e shows a partial view of the AM Results of SNP Allele Combination analysis in PATXS. A total of **1079 significant associations** were obtained for the **2** traits.

Fig. 4.3.5f summarizes the AM Results of both SNP allele analysis and AC effect analyses for AUD and DF8 (Partial view).

Fig. 4.3.5g shows the Model analyses results of AL and AC models for both traits. Significant Correlations between predicted and observed performances are lower than in the PATXF project, particularly for AL models and DF8. They are **59% for DF8** and **75% for AUD,** explaining **34** and **55%** of the total variance, respectively. Notice that the correlations are much higher for the more powerful AC models, since they also consider heterozygous SNP states (**69% for DF8** and **78% for AUD,** explaining **47** and **60%** of the total variance).

Fig. 4.3.5h presents a partial view of the **Breeding values** and **Mean performances** of the genotypes for all traits in **PATXS** if used for breeding and **Fig. 4.3.5i** and **Fig. 4.3.5j** presents a partial view of the corresponding prediction matrixes of progeny performances for crosses between PATXS genotypes for the selected traits derived from AL and AC models, respectively. The user can derive the TOP crosses for each trait by selecting the highest values from this matrix.

We have also determined the Most efficient markers by Multiple regression analyses and extracted the top crosses for AUD and DF8 traits in AL and AC models.

Fig. 4.3.5k shows a reduced set of most efficient markers derived from AL and AC models as obtained by Multiple Regression explaining a large amount (around 90%) of the total variance.

Fig. 4.3.5l presents a partial view of the top crosses for each trait which were extracted from the PPP matrixes of Allele Models and AC models, respectively.

The EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATXS.xlsx**") with all mentioned results can be also accessed in: **PATVIEW>DATABASES>AM-Results>PATXS DB**.

PATRS:

The analogous analyses were also performed for the SNP markers derived from RAD markers and the corresponding figures were produced:

Fig. 4.3.6a contains a partial view of the SNP Characteristics of Project**: PATRS**. A total of **242 nr** SNP markers were available for the analyses.

Fig. 4.3.6b shows partially the molecular patterns of SNP markers of **PATRS** in the Collection,

The Phenotypic Data for **AUD** and **DF8** are here the same as for PATXS.

Fig. 4.3.6d shows a partial view of the AM Results of SNP allele analysis (Allele effects) in PATRS. A total of **308 significant associations** were obtained for the **2** traits.

Unlike in PATXF, AL and AC models are different here, since also heterozygous SNP genotypes can exist for each SNP.

Therefore, Fig. 4.3.6e shows a partial view of the AM Results of SNP Allele Combination analysis in PATRS. A total of **355 significant associations** were obtained for the **2** traits.

Fig. 4.3.6f summarizes the AM Results of both SNP allele analysis and AC effect analyses for AUD and DF8 (Partial view).

Fig. 4.3.6g shows the Model analyses results of AL and AC models for both traits. Significant Correlations between predicted and observed performances are different than in the PATXR project. They are **78% for DF8** and **60% for AUD,** explaining **60** and **35%** of the total variance, respectively. Notice that againthe correlations are much higher for the more powerful AC models, since they also consider heterozygous SNP states (**81% for DF8** and **66% for AUD,** explaining **64** and **43%** of the total variance).

Fig. 4.3.6h presents a partial view of the **Breeding values** and **Mean performances** of the genotypes for all traits in **PATRS** if used for breeding and **Fig. 4.3.6i** and **Fig. 4.3.6j** presents a partial view of the corresponding prediction matrixes of progeny performances for crosses between PATRS genotypes for the selected traits derived from AL and AC models, respectively.

We have also determined the Most efficient markers by Multiple regression analyses and extracted the top crosses for AUD and DF8 traits in AL and AC models.

Fig. 4.3.6k shows a reduced set of most efficient markers derived from AL and AC models as obtained by Multiple Regression explaining a large amount (around 90%) of the total variance.

Fig. 4.3.6l presents a partial view of the top crosses for each trait which were extracted from the PPP matrixes of Allele Models and AC models, respectively.

The EXCEL file ("**E:\VIEWER\PATVIEW\DB\PATRS.xlsx**") with all mentioned results can be also accessed in: **PATVIEW>DATABASES>AM-Results>PATRS DB**.

Cross Predictions

We also analysed the correlations between observed performance and cross predicted Breeding values between Peruvian and Ecuadorian accessions and vice versa for the common traits ATW, NT and Yield. Markers of the projects PATXf, PATRf, PATSF and PATSF were used for this purpose. However the cross predicted correlations were generally low, and in most cases not significant (results not shown). In this sense it was appropriate to make separate analyses for these traits for the Accessions from Peru and Ecuador, since the plant material is very divergent. The accessions from Ecuador consider mainly *S. andigena* materials (and a few native species), while the accessions from Peru are mainly composed of different native species.