Tree Genetics & Genomes (2017) 13: 127
https://doi.org/10.1007/s11295-017-1209-x

ORIGINAL ARTICLE

Assessment of gene copy number variation of Scots pine thaumatin-like protein gene using real-time PCR based methods

Vilnis Šķipars[1] · Elza Rauda[1] · Ilze Snepste[1] · Baiba Krivmane[1] · Dainis Rungis[1]

Received: 9 August 2017 / Revised: 17 October 2017 / Accepted: 8 November 2017 / Published online: 17 November 2017

Abstract The importance and impact of gene copy number variations (CNVs) as a source of polymorphism in the human and other genomes is being increasingly recognized. Less information is available about CNVs in forest tree species, mainly due to the relative lack of genomic resources. In this study, several methods—quantitative polymerase chain reaction, comparative high-resolution melting curve analysis (C-HRM), and digital polymerase chain reaction (dPCR)—were used to investigate CNV of the Scots pine thaumatin-like protein gene (*PsTLP*). The obtained results were supported by transcriptome analysis of a single *Pinus sylvestris* individual and publically available pine genome sequences. Although estimations of gene copy number (CN) varied, depending on the region of the *PsTLP* gene investigated and the endogenous control utilized, our results revealed the existence of copy number variations of the *PsTLP* gene between Scots pine individuals. Of 23 individuals analyzed, two had an increased calculated relative CN regardless of the analyzed gene region and endogenous control used, while several samples had increased copy numbers of regions of the *PsTLP* gene. C-HRM results were highly correlated with qPCR data ($R^2_{TLP3'} = 0.88$; $R^2_{TLPc} = 0.92$), but interpretation of gene CN from C-HRM results proved to be difficult. The results from selected samples analyzed by digital PCR also were highly correlated with qPCR results ($R^2 = 0.90$).

Keywords Copy number variation · *Pinus sylvestris* L. · qPCR · Comparative high resolution melting curve analysis · Thaumatin-like protein · *Heterobasidion annosum*

Communicated by P. Ingvarsson

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11295-017-1209-x) contains supplementary material, which is available to authorized users.

✉ Vilnis Šķipars
vilnis.skipars@silava.lv

[1] Genetic Resource Centre, Latvian State Forest Research Institute "Silava", 111 Rigas Street, Salaspils LV-2169, Latvia

Introduction

Gene copy number variations (CNVs) have been recognized as a major source of variation in humans and other mammals (Iafrate et al. 2004; Sebat et al. 2004; Freeman et al. 2006) as well as in maize (Springer et al. 2009). The duplicated genomic segments leading to CNV are usually reported to be larger than 1 kb (Stankiewicz and Lupski 2010) and can contain one or more genes. CNVs can be classified into two groups based on their frequency in populations—recurrent CNVs and rare CNVs, which are likely to be induced by differing mechanisms. The most common genetic mechanism causing duplication or deletion type recurring CNVs in humans is NAHR (non-allelic homologous recombination) while mechanisms like FoSTeS (Fork Stalling and Template Switching) and MMBIR (microhomology-mediated break-induced replication) are involved in rare CNV events and make use of replication mechanisms (Liu et al. 2012). Non-homologous mechanisms, such as multiple NHEJ (non-homologous end joining), may also account for some complex rearrangements (Gu et al. 2008).

CNVs have not been investigated extensively in conifers, although the amount of genomic information and quality of this information is increasing (Nystedt et al. 2013; Zimin et al. 2017) and initial studies indicate that they may be quite common in spruce (Prunier et al. 2017). Conifer genomes have several properties that may facilitate CNV formation such as a

high proportion of repetitive sequences which can facilitate unequal crossovers and other genomic rearrangements (Gu et al. 2008) as well as the presence of gene family clusters (Liu and Ekramoddoullah 2009; Hedman et al. 2013; Warren et al. 2015) and presence of active transposons (Voronova and Rungis 2014). While the genome segments involved in CNV can be large, the distribution of genes in conifer genomes, averaging in one gene in 705 kb in *Picea abies* (Nystedt et al. 2013), may imply that duplication of large genomic segments could involve only one gene. In the maize genome, the overwhelming majority of CNV events involve one gene (Swanson-Wagner et al. 2010). There is evidence suggesting that gene duplicates from whole genome duplication events diversify developmental and physiological regulation but tandem duplicates increase the diversity in genes involved in environmental response including resistance to pathogens (Salojärvi et al. 2017).

Individuals containing multiple copies of a gene can have higher levels of gene expression, thus influencing the phenotype (Chen et al. 2006; Sutton et al. 2007; Díaz et al. 2012, Mehta et al. 2014). CNV analyses of quantitative trait loci in tree species are scarce, but there are some reports in crop species. Increased copy number of the wheat *Rht-D1b* allele is correlated with yield (Pearce et al. 2011; Li et al. 2012), duplication of the *ZMM19 MADS*-box gene leads to changes in cob phenotype in maize (Wingen et al. 2012), and CNVs influencing growth and development have also been identified in the potato genome (Iovene et al. 2013). In addition, CNVs have been shown to influence pest resistance in soybean (Cook et al. 2012) and glyphosate resistance in *Amaranthus palmeri* (Gaines et al. 2010).

CNVs can be detected using several methods including representational oligonucleotide microarray analysis (ROMA) (Lucito et al. 2003), fosmid paired end sequencing (Tuzun et al. 2005), fluorescent in situ hybridization (FISH), comparative genomic hybridization (CGH) (Kallioniemi et al. 1992; Ju et al. 2010), array comparative genomic hybridization (aCGH) (Perry et al. 2008), use of high-density whole-genome SNP microarrays (Huang et al. 2004), digital PCR (Dube et al. 2008), several next-generation sequencing approaches (Krumm et al. 2012; Duan et al. 2013; Wang et al. 2014; D'Aurizio et al. 2016), and pyrosequencing (Cantsilieris et al. 2013). However, real-time PCR remains the reference method most often used to confirm CNVs identified by other methods (Hashemi et al. 2013; Ghosh et al. 2014).

Scots pine is ecologically and commercially the most important tree species in Latvian forests, being the dominant species in 29% of forests (more than 0.97 million ha) (Ministry of Agriculture of the Republic of Latvia 2014). A breeding program for Scots pine has been established in Latvia, and the infrastructure of this breeding program includes seed orchards and tree nurseries. One of the traits of interest for pine breeding is the resistance to root rot caused by *Heterobasidion annosum*, but this trait has not been included in the breeding program as it is difficult to characterize the degree of resistance against *H. annosum* in Scots pine. Research into the molecular genetic responses of conifers to *Heterobasidion* infection has identified differentially expressed resistance genes (Adomas et al. 2007) as well as differences in expression levels between individuals (Danielsson et al. 2011).

In order to further investigate the basis of this variation, qPCR and C-HRM (Borun et al. 2014) were utilized to analyze CNV of the Scots pine thaumatin-like protein (*PsTLP*) gene. In vitro analyses have shown that the protein encoded by this gene inhibits the growth of *H. annosum* as well as a range of other fungi (Snepste et al., submitted). An initial investigation by qPCR using one primer set revealed evidence of CNV of the *PsTLP* gene in Latvian Scots pine populations (Šķipars et al. 2011). In this study, three primer sets were used to determine the relative amplicon quantities of different regions of the *PsTLP* gene using qPCR, and two primer sets were used for C-HRM analysis. Three endogenous control genes were used in qPCR. In addition, a limited number of samples were also analyzed using digital PCR (dPCR). Usually, detection of CNV using qPCR utilizes reference samples with predetermined copy number (D'haene et al. 2010). However, there are no Scots pine reference samples with well-characterized gene copy numbers that could be utilized. Additional evidence of existence of multiple copy numbers of the *TLP* gene were obtained by analysis of *Pinus sylvestris* transcriptome data obtained from a single individual and publically available genomic sequence scaffolds of *Pinus taeda* and *Pinus lambertiana*.

Materials and methods

Experimental material

Twenty-three mature Scots pine individuals (GE05, GE06, and GE09–GE29) were utilized for CNV analyses of the *PsTLP* gene using the qPCR and C-HRM methods. The trees originate from a pine breeding program progeny trial established in 1979 located in Kalsnava district, Latvia. Samples are open-pollinated progeny obtained from a number of different mother trees, and are a sub-set of samples previously analyzed for CNVs (Šķipars et al., 2011). DNA was extracted from fresh needles using the Genomic DNA isolation kit (Thermo Fisher Scientific) and quantified by use of Qubit fluorometer and the dsDNA BR kit (Thermo Fisher Scientific). The integrity of DNA was assessed by electrophoresis on a 2% agarose gel.

qPCR

CNV of the *PsTLP* gene (GenBank accession no. JX461338.1, total length 936 bp, CDS 43-121, 267-892) was analyzed using three different primer sets, amplifying separate, non-overlapping regions of the *PsTLP* gene. Primer set *TLP3′* amplifies the region from nucleotide 797 to 861 of the *PsTLP* gene, primer set *TLPc* amplifies the region from nucleotide 474 to 693, and primer set *TLP5′* amplifies the region from nucleotide 62 to 161. Amplicons of primer sets *TLP3′* and *TLPc* are entirely within the protein coding region of the gene, while the amplicon of primer pair *TLP5′* contains both protein coding and intron sequence. The amplicon of *TLPc* partly covers the sequence encoding the signature amino acids of the thaumatin family (Prosite accession no. PS00316) (Fig. 1).

All *PsTLP* specific primer sets were analyzed with three different endogenous controls—glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), an *Avr9*/Cf-*9* rapidly elicited (*ACRE*) gene homolog (*PsACRE*), and a *Pinus taeda* water-stress inducible protein (*Lp3–1*) gene. The *GAPDH* gene is widely utilized as control gene for quantification of gene expression; therefore, the utility of this gene as a control for gene copy number was investigated. The *PsACRE* (an *Avr9*/Cf-*9* rapidly elicited (*ACRE*) gene homolog was chosen as a control as in previous experiments in our laboratory; it had constant amplification results from DNA samples. The *Pinus taeda* water-stress-inducible protein (*Lp3-1*) gene (NCBI accession number U52865) has been characterized as a conserved ortholog sequence in conifers (Krutovsky et al. 2006), and therefore was expected to be stable and conserved within *P. sylvestris*. All primers were screened for specificity in silico using NCBI BLAST (Altschul et al. 1990), and no significant similarities with other sequences were found. The utilized primer sequences are given in Table 1.

The qPCR protocol for determination of relative CN of the *PsTLP* gene (reaction volume 10 μl) was as follows: 2 μl of 5× HOT FIREPol® EvaGreen® HRM Mix (Solis BioDyne), 250 nM each forward and reverse primer, 5 ng of Scots pine DNA, deionized water. Thermal cycling conditions were as follows: 15′ 95 °C initial denaturation and polymerase activation followed by 40 cycles of 95 °C 15 s, 60 ° C 20 s, 72 °C 1 min. Data interpretation is described in D'haene et al. (2010) and Šķipars et al. (2011).

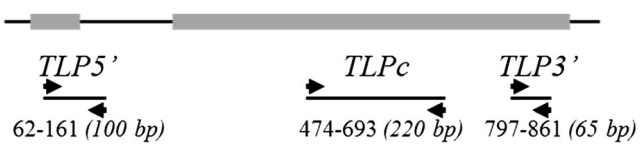

Fig. 1 Schematic depiction of the *PsTLP* gene. Coding regions are presented by gray bars; 5′ and 3′UTRs and the intron are represented as a black line. Regions amplified by primer pairs used in this study are depicted below

For valid ΔΔCT calculations, the amplification efficiencies of target amplicons (*TLP3′*, *TLPc*, *TLP5′*) and endogenous control amplicons (*GAPDH*, *PsACRE*, *Lp3-1*) must be within 10% of each other (Schmittgen and Livak 2008). Amplification efficiency was determined by the CT slope method. CT values were measured over a twofold dilution range (1.75–14 ng) of three DNA samples (individuals GE05, GE14, GE15). The amplification efficiencies were as follows: *TLP5′*, 100.41%; *TLPc*, 96.93%; *TLP3′*, 96.22%; *GAPDH*, 102.32%; *PsACRE*, 98.07%; *Lp3-1*, 95.62%. In previous reports, standard samples with known target gene copy number were utilized for calculation of the rescaling factor. However, such standard samples are not available for Scots pine; therefore, samples utilized for rescaling factor calculations were chosen from among the analyzed samples. As it is expected that the majority of samples would show similar results (representing the most common gene copy number), the samples belonging to the majority (by relative quantitation results) could be used as reference samples. The rescaling factor is utilized to classify quantitative results into discrete relative gene copy number classes, and therefore does not influence the relative ranking of the individuals, but may affect the interpretation of the gene copy number for individuals with quantitation results close to the boundaries between relative gene copy number classes. Samples used for rescaling factor calculation and the calculated rescaling factors are given in Table 2.

C-HRM

Due to the multiplex nature of the C-HRM reaction, the *GAPDH* and *PsACRE* genes, which were previously used as endogenous controls in qPCR analysis, were unable to be used as controls because of overlapping melting temperatures of amplicons of these control genes with the amplicon of interest. Therefore, only the *Lp3-1* gene was used as the endogenous control for C-HRM analysis. Efficiency of the multiplex C-HRM reaction was tested by determining the PCR efficiency for individual primer sets and for the multiplex reaction using fivefold serial dilutions (2.4 to 300 ng per reaction). qPCR reactions with 2.4 to 60 ng of DNA per reaction showed no sign of significant PCR inhibition, in contrast to reactions with 300 ng of DNA per reaction (*TLP3′* 110.98%, *Lp3-1* 91.57%, multiplex 100.95%). C-HRM efficiency of the reaction with *TLPc* primer set was determined by analyzing the results of a twofold serial dilution (2.5–20 ng), and *TLPc* amplification efficiency was 104.22% (multiplex 94.41%). Fifteen nanograms of DNA per reaction were utilized for C-HRM analyses. Primer set *TLP5′* was not used in C-HRM analysis due to overlapping melting temperature of the amplicon with control amplicons. Reaction conditions (total volume 20 μl) were as follows: 4 μl of 5× HOT FIREPol® EvaGreen® HRM Mix (Solis BioDyne), 250 nM each forward and reverse primer,

Table 1 Sequences of PCR primers utilized in this study

Primer set	Forward/reverse/probe[a]	Sequence (5′ → 3′)
TLP5′	F	CAGGGTCCCTTTGGATCAC
	R	ATAGTGATATTGTAGAGTAATTGAGAGAGC
TLPc	F	GTGGTGGGTTGCTCAATTGTC
	R	CCATCGGTCACTTTCAGTTCTG
TLP3′	F	CAGTGCCCACAGGCATACAG
	R	CCACCAGGGCAGGTGAAG
	P	6FAM-TATGCCAAGGACGATGCCACCAGC-TAMRA
GAPDH	F	ACGGTTTTGGTCGAATTGGA
	R	CCCCACGAGCTCGATATCAT
	P	VIC-CTCGTCGCCCGTGTGGCTCTG-TAMRA
PsACRE	F	CATCATTACTTCCCCACACATATTCT
	R	TGGGCTCTTCCTTGTCTTCAA
Lp3–1	F	TCTGGCTGGACACATCATGAA
	R	GAGGGACTAATAACCCGTGATGATA

[a] For use in dPCR

15 ng of Scots pine DNA, deionized water. Thermal cycling conditions: 15′ 95 ° C initial denaturation and polymerase activation followed by 26 cycles of 95 °C 15 s, 60 °C 20s, 72 °C 20 s followed by a high-resolution melting curve stage. The reaction was performed on an Applied Biosystems StepOnePlus instrument.

In the original report about this method, standard samples with known copy number of the target genes were used to obtain the peak height ratio for use in data normalization, assuming that the peak height ratio is 0.5 × for samples with a gene deletion (n copies) and 1.5 × for samples with a duplication (3n copies) (compared to the peak height ratio of the standard samples) (Borun et al. 2014). Data interpretation involves data normalization which essentially means that the peak height ratio is divided by the average peak ratio for the standard samples (Borun et al. 2014). After including standard deviations, an approximate scale was created by the authors of this method for assignment of analyzed samples to different sample groups. Samples with a normalized peak height ratio below 0.6 indicate a deletion, a ratio between 0.9 and 1.1 indicates unchanged copy number compared to controls, and a ratio above 1.4 indicates duplication (Borun et al. 2014). We extrapolated this scale so a value of 2.0 ± 0.1 would correspond to relative copy number of 4n.

Digital PCR

Digital PCR was performed on the Life Technologies QuantStudio® 3D Digital PCR System using QuantStudio™ 3D Digital PCR Master Mix, and data were analyzed using QuantStudio® 3D AnalysisSuite™ Cloud Software.

The composition of one reaction with total volume of 15 μl was 7.5 μl of Digital PCR Master Mix, *PsTLP* assay containing primers *TLP3′-F* and *TLP3′-R* with final concentration 900 nM and probe *TLP3′-P* with final concentration 300 nM, *GAPDH* assay containing primers *GAPDH-F* and *GAPDH-R* with final concentration 900 nM and probe *GAPDH-P* with final concentration 300 nM, 20 ng of genomic DNA. Each reaction (14.5 μl) was loaded onto a QuantStudio™ 3D Digital PCR 20K chip. PCR was performed on a GeneAmp

Table 2 Samples used for calculation of rescaling factors

Endogenous control	Assay	Samples used for calculation of rescaling factors	Calculated rescaling factor
GAPDH	*TLP5′*	GE05, GE09, GE16, GE24, GE28	0.492
	TLPc	GE05, GE13, GE16, GE23, GE26	0.514
	TLP3′	GE05, GE11, GE16, GE24, GE28	0.498
PsACRE	*TLP5′*	GE05, GE09, GE11, GE12, GE16	0.398
	TLPc	GE05, GE06, GE11, GE23, GE24	0.582
	TLP3′	GE05, GE10, GE16, GE20, GE28	0.504
Lp3-1	*TLP5′*	GE05, GE09, GE16, GE24, GE28	0.398
	TLPc	GE05, GE09, GE10, GE24, GE28	0.646
	TLP3′	GE05, GE10, GE18, GE23, GE25	0.512

PCR System 9700. The cycling conditions were 10 min at 96.0 °C followed by 39 cycles of 2 min at 60.0 °C and 30 s at 98.0 °C, then a hold for 2 min at 60.0 °C followed by storage at 10.0 °C in the instrument until the reading of the chips.

Transcriptome sequencing

Transcribed sequences were obtained from analysis of one clone (sample GE24) after inoculation with *H. annosum* (strain V Str 28). RNA was extracted following the method described in Šķipars et al. (2014); obtained RIN (RNA integrity number) values exceeded 7. Ribosomal RNA was removed using the Thermo RiboMinus™ Plant kit for RNA-Seq, and the transcriptome libraries were prepared using the Ion Total RNA-Seq Kit v2 (both kits from Thermo Fisher Scientific). The following steps including emulsion PCR and IonTorrent sequencing were performed at the Latvian Biomedical Research and Study Centre. Transcriptome reads were aligned against expected amplicon sequences for primer sets *TLP5′*, *TLPc*, and *TLP3′*, alignment limited to 100 best hits. For graphical depiction, transcriptome sequences were trimmed and aligned to the amplicons; singleton sequences were removed. Sequences were grouped into haplotypes manually. The transcriptome read database used for the analysis contained 60 million reads. Software analyses were performed using CLC Genomics Workbench (Qiagen) and Vector NTI (InforMax Inc.).

Results

Analysis of the qPCR results (supplementary Table 1) indicated that both the gene region amplified and the endogenous control used in the analysis can have an effect on estimated relative copy number of the *PsTLP* gene (Table 3). Comparison of the calculated relative gene copy number interpretation results revealed that, depending on the endogenous control used, six to seven samples had the same copy number of all three *PsTLP* gene regions. There are four samples (GE06, GE10, GE16, and GE21) which had the same copy number for each gene region regardless of the endogenous control utilized. Two samples, GE16 and GE21, had the same calculated relative gene copy number for all three gene regions with all endogenous controls. Sample GE19 had the same region specific calculated relative gene copy number with all endogenous controls. In four cases, the relative copy number of the *TLP3′* region was increased by two copies or more compared to the estimated copy numbers of the other two gene regions. These samples include GE09, GE17, GE27, and GE29, regardless of the endogenous control used. In contrast, in sample GE18, the copy number of the *TLP3′* region decreased by two copies compared to copy numbers of the other gene regions, regardless of the endogenous control utilized. This indicates that the 3′ region was more variable in terms of copy number in comparison to the 5′ and central regions, regardless of the endogenous control utilized. The calculated relative copy number of the *TLPc* region for sample GE22 increased by two copies when *GAPDH* was used as the endogenous control and a decreased copy number for sample GE14 was calculated when *Lp3-1* was used as the endogenous control. The copy number of the *TLP5′* region did not have a difference of more than two copies between the analyzed individuals, regardless of the endogenous control. Samples GE06, GE10, and GE13 had an endogenous control—specific increase or decrease of estimated gene copy number. These results highlight the necessity of using several gene regions and several endogenous controls to ensure accurate CNV assay results.

All qPCR analyses were performed using 5 ng of template DNA. This provides the opportunity to not only use the relative quantity values determined by use of reference samples and endogenous controls but to also analyze the raw Ct values, which are expected to be very similar between samples for the endogenous controls as well as the target amplicons for samples with similar gene region copy numbers. Analysis of the deviation of sample Ct values from average Ct values for endogenous control amplicons reveal differences in Ct values between individuals (Fig. 2).

The observed deviations can be expressed as the theoretical influence of the deviation in the endogenous control reaction on the relative quantitation results (supplementary Table 2). Examination of the relative gene copy number in conjunction with the information about raw Ct values and the deviations of the Ct values from the average allows assessment of whether the change in calculated relative quantity of target amplicon is due to amplification of the target amplicon or to unexpected variation in amplification of the endogenous control, which would indicate that the change in relative gene copy number may be artefactual. For example, the increase in the relative gene copy number in sample GE6 (using *Lp3-1* as the endogenous control) is probably artefactual due to the anomalous amplification of the *Lp3-1* endogenous control in this individual (Fig. 3). However, use of this information for correction of relative quantification results and interpretation of relative gene copy number is complicated by possible variations in qPCR efficiency and technical replicate Ct values as well as by the fact that the average Ct value is calculated from all samples, including those with deviations. Therefore, this information can be utilized as an indicator to identify possibly anomalous results or samples, which should be further investigated with regard to CNV of the target gene or gene region.

To confirm these results using an alternative CNV detection technique, the 23 individuals were analyzed using C-HRM. The multiplex C-HRM reaction produces two amplicons with distinct melting temperatures in a single

Table 3 Comparison of calculated relative copy number values by qPCR of the three regions of the *PsTLP* gene using three endogenous controls

Endogenous control	*GAPDH*			*PsACRE*			*Lp3-1*		
Region / Sample	*TLP5′*	*TLPc*	*TLP3′*	*TLP5′*	*TLPc*	*TLP3′*	*TLP5′*	*TLPc*	*TLP3′*
GE05*	2	2	2	3	2	2	3	2	2
GE06	2	2	2	2	2	2	6	6	6
GE09	2	2	5	2	2	4	2	2	4
GE10	1	1	1	2	2	2	2	2	2
GE11	2	2	2	2	2	2	2	2	1
GE12	3	3	4	2	3	4	3	3	5
GE13	2	2	2	1	2	3	1	1	1
GE14	3	4	4	3	3	4	7	4	6
GE15	3	3	2	3	3	2	3	2	1
GE16	2	2	2	2	2	2	2	2	2
GE17	2	2	5	2	2	5	3	3	6
GE18	4	5	2	4	4	2	4	5	2
GE19	3	3	4	3	3	4	3	3	4
GE20	4	5	3	3	3	2	3	3	2
GE21	2	2	2	2	2	2	2	2	2
GE22	3	5	2	3	4	2	3	3	2
GE23	1	2	1	2	2	1	2	3	2
GE24	2	3	2	2	2	2	2	2	2
GE25	2	3	2	2	2	1	3	3	2
GE26	1	2	2	1	2	2	1	1	1
GE27	4	5	10	3	3	6	2	3	7
GE28	2	3	2	2	3	2	2	3	3
GE29	2	3	7	1	2	7	2	2	6
mean	2.35	2.87	3.04	2.26	2.48	2.83	2.74	2.70	3.09

[a] Reference sample. Green highlight indicates a copy number interpretation of 2 in all assays with the same endogenous control. Red highlight indicates a copy number interpretation exceeding 2 in all assays with the same endogenous control. Yellow highlight indicates a copy number interpretation of 1 in all assays with the same endogenous control

reaction. Of the three previously analyzed *PsTLP* primer sets, only the *TLP3′* and *TLPc* primer sets were compatible with the *Lp3-1*endogenous control for C-HRM analysis. Differences in the copy number of an amplicon are detected by calculating the peak height ratio of the target amplicon and the control amplicon and comparing these values between samples (Fig. 4, Table 4).

As no reference samples were available, we calculated the normalization factor from the peak height ratio values obtained in our experiments. The distribution of peak height ratio results was estimated. Samples were divided into peak height ratio groups by increments of 0.1. As we expect most of the samples to have the same gene copy number, the average value of samples from the largest groups were used as standards. For the *TLP3′* primer set, the normalization factor was calculated to be ~0.796 (the average values of samples with peak height ratio within the range 0.6–0.9), but for the *TLPc* primer set, it was calculated to be ~2.076 (the average values of samples with peak height ratio within the range 2.0–2.2). The distribution of raw peak ratio results is shown in Fig. 5.

Interpretation of the C-HRM results is problematic because many samples are outside the predefined boundaries for

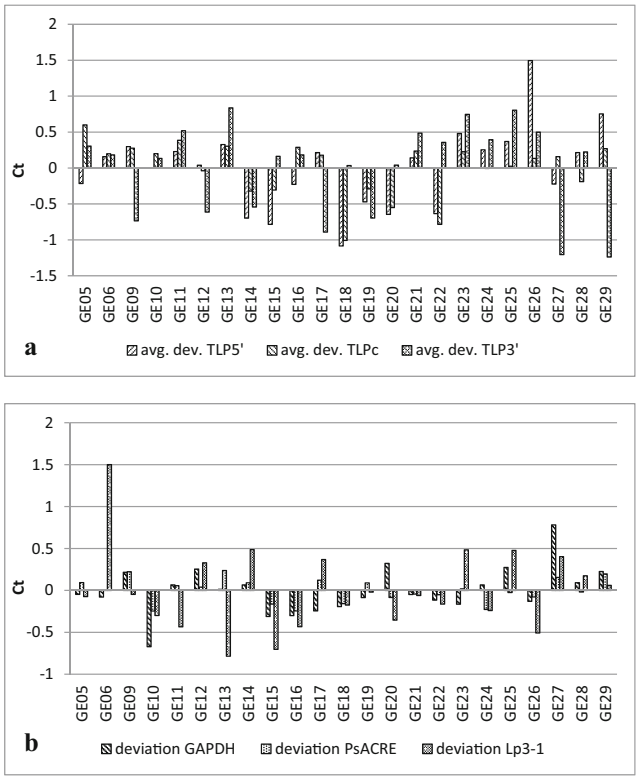

Fig. 2 **a** Comparison of deviations from average Ct value for each sample depending on target region. **b** Comparison of deviations from average Ct value for each sample depending on endogenous control

segregation of the samples into different gene copy number groups. Six of 23 samples analyzed with *TLP3′* primer set and 12 of 23 samples analyzed with primer set *TLPc* fall outside of these boundaries. Nevertheless, it is possible to use the C-HRM results for visualization of differences between samples even if there are some issues regarding gene copy number interpretation (Fig. 6).

The quantitation results obtained by C-HRM are highly correlated with the qPCR results ($R^2_{TLP3'} = 0.88$; $R^2_{TLPc} = 0.92$). As mentioned previously, the interpretation of the quantitation results and assignment of samples into discrete gene copy number groups was complicated by the absence of reference samples with a pre-determined gene copy number. Not all of the samples estimated to have increased copy number by qPCR were confirmed by the C-HRM method, indicating that while the quantitation results were well correlated, the interpretation of gene copy number, particularly in the absence of well-defined reference samples, is more uncertain using the C-HRM method in comparison to qPCR. In addition, the C-HRM method is not as widely applicable to all target gene/endogenous control combinations due to the required differences in amplicon melting temperatures.

A limited number of individuals with differing CN as determined by the qPCR and C-HRM analysis were also analyzed using digital PCR (dPCR) (Table 5). Despite the fact that the dPCR method results in absolute numbers of an amplicon (copies/μl which can be translated in copies/genome if the mass of DNA per genome is known), a reference gene (in this instance *GAPDH*) was included in the analysis for normalization. While the absolute calculated amplicon numbers of the *PsTLP* 3' region were different, the relative values determined by dPCR were correlated with the qPCR data for the analyzed samples ($R^2 = 0.90$) (which were also normalized using the *GAPDH* control gene). However, the sample number is too small for any meaningful conclusions to be made, and these data should be viewed only as additional supporting information.

Transcriptome data were also used to investigate the copy number of the analyzed *PsTLP* gene regions within the genome of one individual (GE24). This individual showed similar quantitative and calculated relative copy number results for all three *PsTLP* gene regions. The transcript sequences were aligned to the sequences of amplicons produced by the primer sets *TLP5′*, *TLPc* and *TLP3′* (supplementary Fig. 1). BLAST results were limited to 100 reads most similar to the

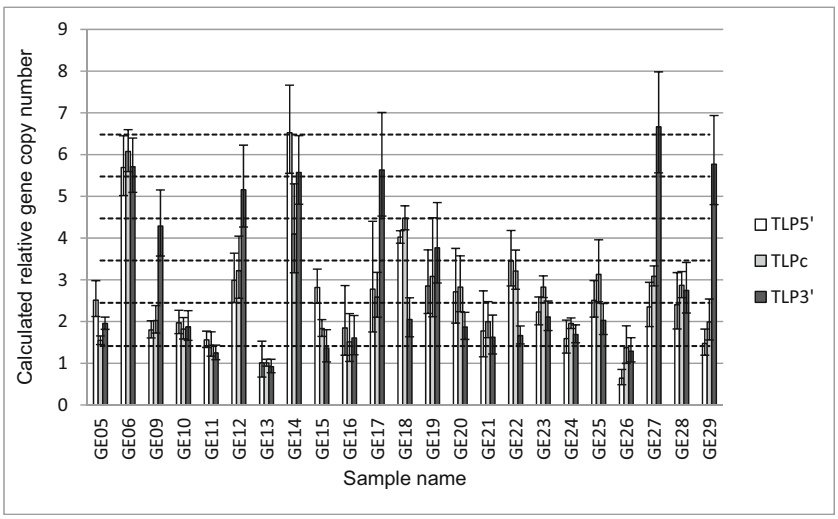

Fig. 3 Relative gene copy number of three different *PsTLP* gene regions, endogenous control *Lp3-1*; dashed lines represent borders between relative gene copy number groups

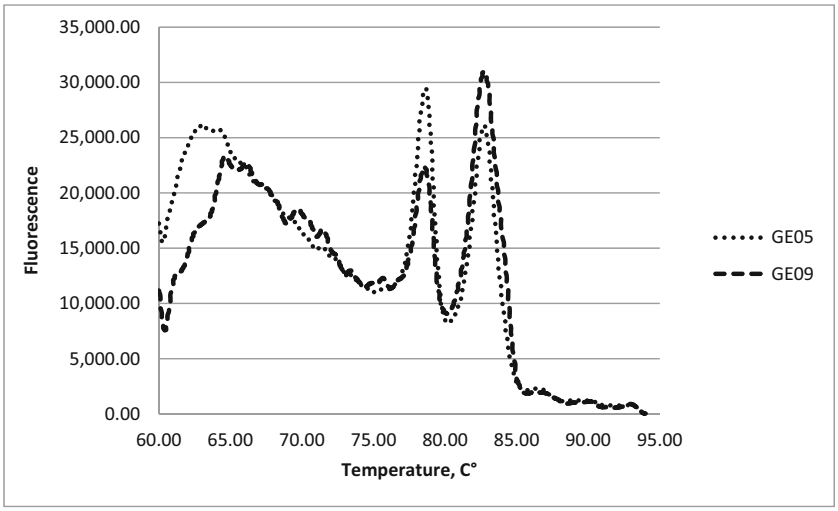

Fig. 4 Derivative melting curves of C-HRM analyses of samples GE05 and GE09. The left peaks (between 75 and 80 °C) are the reference amplicon (*Lp3-1*) melting curve peaks and the right peaks (between 80 and 85 °C) are the *PsTLP* amplicon melting curve peaks (primer set *TLP3′*)

amplicon sequences. For the *TLP5′* region, the expressed sequence reads only mapped to the exon. After unique (singleton) reads were removed, 14 SNPs were identified in the *TLP5′* region (corresponding to 4 haplotypes), 30 SNPs were identified in the *TLPc* region (corresponding to 5 haplotypes), and 20 SNPs were identified in the *TLP3′* region (corresponding to 8 haplotypes). These results indicate that there are several variants of the *PsTLP* gene within the genome of individual GE24, with a differing number of haplotypes of each amplified gene region. While no copy number differences between the three analyzed gene regions were identified in this individual, the identification of multiple haplotypes found in the gene transcripts supports the presence of multiple copy numbers of all or part of the *PsTLP* gene. In addition, a

Table 4 C-HRM results of *PsTLP* assays

Sample	C-HRM result (*TLP3′*)	Normalized C-HRM result (*TLP3′*)	C-HRM interpretation (*TLP3′*)	C-HRM result (*TLPc*)	Normalized C-HRM result (*TLPc*)	C-HRM interpretation (*TLPc*)
GE05	0.86	1.08	2n	1.99	0.96	2n
GE06	1.65	2.07	4n	3.91	1.88	3n
GE09	1.39	1.74	3n	2.04	0.98	2n
GE10	0.87	1.09	2n	2.09	1.01	2n
GE11	0.68	0.86	–	1.76	0.85	–
GE12	1.47	1.85	3n	2.60	1.25	–
GE13	0.61	0.77	–	1.39	0.67	–
GE14	1.51	1.90	4n	3.37	1.62	–
GE15	0.71	0.89	–	2.39	1.15	–
GE16	0.79	0.99	2n	2.09	1.01	2n
GE17	1.53	1.92	4n	2.49	1.20	–
GE18	0.77	0.96	2n	3.38	1.63	3n
GE19	1.18	1.48	3n	2.86	1.38	–
GE20	0.87	1.09	2n	2.77	1.34	–
GE21	0.87	1.09	2n	2.26	1.09	2n
GE22	0.83	1.04	2n	3.22	1.55	3n
GE23	1.00	1.26	–	2.66	1.28	–
GE24	0.84	1.05	2n	2.28	1.10	2n
GE25	0.88	1.11	–	2.89	1.39	–
GE26	0.77	0.97	2n	1.88	0.90	2n
GE27	1.52	1.91	4n	2.74	1.32	–
GE28	1.01	1.27	–	2.70	1.30	–
GE29	1.63	2.05	4n	2.18	1.05	2n

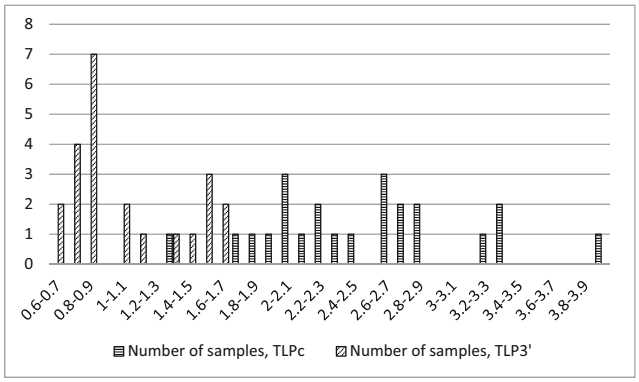

Fig. 5 Distribution of raw peak ratio results by target region

larger number of transcribed sequence variants were found corresponding to the 3′ region of the *PsTLP* gene, suggesting a higher copy number of this region of the gene, indicating that different regions of the *PsTLP* gene may have differing copy numbers, which corresponds to the overall results obtained by real-time PCR.

The *PsTLP* gene sequence was also used to search the *Pinus taeda* genome (NCBI accession no. APFE000000000.3) sequence scaffolds using NCBI BLAST. More than one match was found to several scaffolds—130,911 (2 hits), 85,527 (3 hits), 51,749 (3 hits) (accession numbers APFE031015264.1, APFE030842585.1, and APFE031073227.1, respectively). In many cases, these matches are missing the 5′ part of the gene and the aligned sequence starts after the intron (all matches from scaffolds 85,527 and 51,749) and one of the matches to scaffold 51,749 contains only a 197 nt long sequence from the 3′ region of the gene. Similarly, BLAST analysis of the whole genome shotgun sequencing project of *Pinus lambertiana* (NCBI accession no. LMTP000000000.1) using the full-length *PsTLP* gene as the query sequence identified six scaffolds with more than one match to the query sequence. One matching sequence contained the entire *TLP* gene; three hits to different scaffolds contained the central and the 5′ regions of the gene, while other hits included only one of the gene regions. Five hits from three different scaffolds contained the intron sequence. In this description, we use "5′ region, central region, and 3′ region" to describe whether the matching sequences contain the sequences of the amplicons generated with our primer sets. Detailed alignment information is provided in supplementary Table 3.

Discussion

Results of this study show that to reliably detect gene CNV by quantitative PCR methods, it is necessary to use several primer sets targeted to different regions of the target gene, as demonstrated by the detected differences in relative quantity between different regions of *PsTLP*. The *PsTLP* gene copy number results obtained using the different methods used in this study are comparable, best demonstrated by the strong correlation of raw data values. Given that the PCR-based methods utilized the same primer sets, this is not unexpected. However, there were quantitative differences between the three analyzed gene regions within some individuals suggesting presence of partial duplications of the *Ps TLP* gene. In addition, the endogenous control genes utilized also had an influence on the calculated relative gene copy number results, indicating that several control genes should be utilized, in order to detect false positive gene CNV results. Comparing the PCR-based methods utilized in this study, C-HRM is more limited regarding experimental design compared to qPCR, as target and control amplicons are multiplexed and therefore are required to have differing melting temperatures. In addition, reference samples with defined target gene copy numbers are required for accurate gene copy number interpretation. The digital PCR

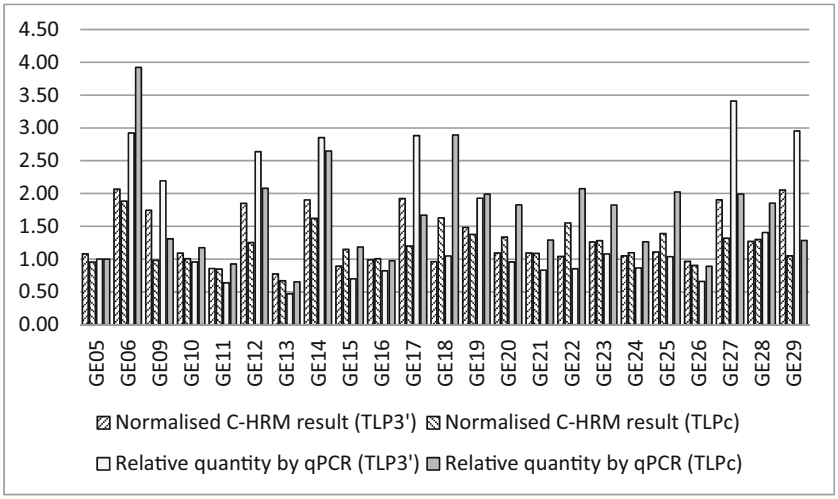

Fig. 6 Comparison of normalized C-HRM results with relative quantification results obtained by use of qPCR. In both methods primer sets *TLP3′* and *TLPc* were used, endogenous control *Lp3-1*

Table 5 Results of CNV analysis using dPCR

Sample name	Calculated relative *PsTLP* copy number by qPCR (primer set *TLP3′*)	*PsTLP* copy number by dPCR (primer set *TLP3′*)
GE27	10	0.913
GE26	2	0.416
GE13	2	0.493
GE19	4	0.676

results, which can be utilized for determination of absolute gene copy numbers, can also present difficulties in interpretation if the structure of CNV polymorphism is complicated (e.g., a small fold change) without the use of pre-characterized reference samples. A single region of a gene can be utilized to identify CNVs when the gene and surrounding regions have been well characterized by sequencing or other approaches (Anhuf et al. 2003; Kulka et al. 2006; Díaz et al. 2012; Cook et al. 2014). One of the advantages of using CGH for detection of gene CNVs is that thousands of probes per array can be utilized, and the probe design process can include criteria such as the minimum number of probes per gene and distance between probes (Swanson-Wagner et al. 2010; Prunier et al., 2017). Yet, in the interpretation of CGH results, variable signal intensity ratios from different probes from a single gene are often used as criteria for omitting a gene from analysis (Swanson-Wagner et al. 2010; Prunier et al. 2017) without considering possible partial duplication. An alternative CGH-based approach where CNVs are detected based on assessment of signal ratios of adjacent probes (Springer et al. 2009) could be more suited for identification of partial gene duplications. High-throughput sequencing (HTS) methods could provide an alternative approach to study CNV than qPCR or CGH as more information is obtained about possible structural variations (SVs). In addition, this method would not be influenced by SNPs in primer binding sites of a qPCR assay (SNPs in primer binding sites were observed in our analysis of transcriptome data). However, the short read lengths that are a feature of the majority of current HTS technologies complicate the analysis of complex genomic SVs (including CNVs), even in well-characterized genomes (Sudmant et al. 2015). The increasing availability of long read sequencing technologies will simplify the identification and characterization of these complex SVs; however, high-quality reference genomes will still be required to provide accurate genotyping of these SVs and their functional significance (Couldrey et al. 2017).

The significance of gene CNVs is likely to have been underestimated due to the technical difficulties of accurate detection and determination of polymorphisms within populations. Duplicated genome regions have been implicated in the formation of gene families and pseudogenes (Zhang 2003); however, these have been studied after sequence divergence of the duplicated regions. The functions of full-length gene duplicates are retained at a comparatively high frequency, suggesting that positive selection, via several mechanisms, can reduce the rate of pseudogene formation (Moore and Purugganan 2005; Panchy et al. 2016). Most CNV studies have emphasized the detection of duplicated full length genes; however, partial gene duplications can also have a functional role. Partially duplicated genes have been shown to contribute to formation of new genes, frequently with altered or novel functions (Toll-Riera et al. 2011). Examples include the *HvARM1* gene from *Hordeum vulgare*, which contributes to resistance against the powdery mildew fungus *Blumeria graminis* (Rajaraman et al. 2017). In addition, a partial duplication of the *A17* protein encoding gene in vaccinia virus provides resistance to rifampin (Erlandson et al. 2014), while partial duplication of the *COL1A2* gene (Raff et al. 2000) and other genes (Hu and Worton 1992) can cause disease in humans. In addition to qPCR evidence for full or partial duplications of the *PsTLP* gene, analysis of the transcriptome obtained from a single individual identified a number of haplotypes, suggesting that several different *PsTLP* copies are transcriptionally active. Divergence of gene expression between duplicated genes has been reported, and the degree of divergence depends on the mechanisms by which these duplications were formed and the time since duplication (Wang et al. 2011).

The structure and evolution of CNV polymorphism of the *PsTLP* gene are not clear, but an ancestral gene duplication event can be proposed (present in *P. taeda* and *P. lambertiana*), with additional duplications in *P. sylvestris* resulting in the observed differences between *P. sylvestris* individuals. Whole genome duplication events are proposed to have occurred at least two times in the evolution of major gymnosperm clades (Li et al. 2015); thus, these events might have contributed to the observed CN variations. Copy number variations or structural variations in resistance- or stress response-linked genes were found to be common in other studies (Neiman et al. 2009; DeBolt 2010; McHale et al. 2012; Boocock et al. 2015; Prunier et al. 2017). Additional copies of resistance linked sequences should increase the rate of formation of new resistance-linked genes as it would increase the amount of sequences available for homologous recombination linked CNV events. The observed enrichment of defense/immunity related CNVs among the entire set of CNVs identified in *Picea* species (Prunier et al. 2017) might suggest positive selection effects. The formation of resistance gene clusters can generate and maintain high haplotypic diversity, thus facilitating rapid evolution of novel resistance genes (Friedman and Baker 2007). This may indicate that resistance related genes are preferentially duplicated and have a higher frequency of CNVs. Recent studies on CNV mechanisms in different species show that a portion of CNV events

are recurrent and occur at specific genomic locations (Zmienko et al. 2016). Analysis of the genomic regions surrounding these duplications may identify the presence of particular sequence motifs that may be implicated in CNV formation.

In conclusion, the real-time PCR-based methods utilized in this study identified reproducible quantitative differences in the copy number of all or part of the *PsTLP* gene. Our results indicate that two of 23 samples (8.7%) have increased relative copy number regardless of target region and endogenous control, and other individuals have increased copy numbers of regions of the *PsTLP* gene. While in some cases this interpretation could be a result of technical variations in the utilized methods, transcriptome and genome alignment analyses provide additional evidence of partial duplications of the *TLP* gene in conifers. Further analysis of the genomic regions surrounding the *PsTLP* loci in *P. sylvestris* will enable a more thorough characterization of the CNV events and provide insight into the evolution of these events and explain the observed differences between *P. sylvestris* individuals, including transcription profiling of different *PsTLP* transcripts and linking of these data to phenotype.

Acknowledgments This study was funded by the Latvian Council of Science project "Investigation of molecular defense mechanisms in Scots pine (*Pinus sylvestris* L.)" (No. 284/2012). We would like to thank Carl Gunnar Fossdal and Adam Vivian-Smith from Norwegian Institute of Bioeconomy Research for digital PCR.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Data archiving statement DNA sequences are available in NCBI, and their accession numbers are mentioned in text of the manuscript. The haplotype sequences obtained from the transcriptome have been attached as supplementary text file 1 in FASTA format.

References

Adomas A, Heller G, Li G, Olson Å, Chu T-M, Osborne J, Craig D, Van Zyl L, Wolfinger R, Sederoff R, Dean RA, Stenlid J, Finlay R, Asiegbu FO (2007) Transcript profiling of a conifer pathosystem: response of *Pinus sylvestris* root tissues to pathogen (*Heterobasidion annosum*) invasion. Tree Physiol 27(10):1441–1458. https://doi.org/10.1093/treephys/27.10.1441

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anhuf D, Eggermann T, Rudnik-Schöneborn S, Zerres K (2003) Determination of SMN1 and SMN2 copy number using TaqMan technology. Hum Mutat 22(1):74–78. https://doi.org/10.1002/humu.10221

Boocock J, Chagné D, Merriman TR, Black MA (2015) The distribution and impact of common copy-number variation in the genome of the domesticated apple, *Malus* x *domestica* Borkh. BMC Genomics 16(1):848. https://doi.org/10.1186/s12864-015-2096-x

Borun P, Kubaszewski L, Banasiewicz T, Walkowiak J, Skrzypczak-Zielinska M, Kaczmarek Rys M, Plawski A (2014) Comparative-high resolution melting: a novel method of simultaneous screening for small mutations and copy number variations. Hum Genet 133(5):535–545. https://doi.org/10.1007/s00439-013-1393-1

Cantsilieris S, Baird PN, White SJ (2013) Molecular methods for genotyping complex copy number polymorphisms. Genomics 101(2):86–93. https://doi.org/10.1016/j.ygeno.2012.10.004

Chen QR, Bilke S, Wei JS, Greer BT, Steinberg SM, Westermann F, Schwab M, Khan J (2006) Increased WSB1 copy number correlates with its over-expression which associates with increased survival in neuroblastoma. Gene Chromosome Canc 45(9):856–862. https://doi.org/10.1002/gcc.20349

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338(6111):1206–1209. https://doi.org/10.1126/science.1228746

Cook DE, Bayless AM, Wang K, Guo X, Song Q, Jiang J, Bent AF (2014) Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. Plant Physiol 165(2):630–647. https://doi.org/10.1104/pp.114.235952

Couldrey C, Keehan M, Johnson T, Tiplady K, Winkelman A, Littlejohn MD, Scott A, Kemper KE, Hayes B, Davis SR, Spelman RJ (2017) Detection and assessment of copy number variation using PacBio long read and Illumina sequencing in New Zealand dairy cattle. J Dairy Sci 100(7):5472–5478. https://doi.org/10.3168/jds.2016-12199

D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A (2016) Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res 44(20):e154. https://doi.org/10.1093/nar/gkw695

D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50(4):262–270. https://doi.org/10.1016/j.ymeth.2009.12.007

Danielsson M, Lundén K, Elfstrand M, Hu J, Zhao T, Arnerup J, Ihrmark K, Swedjemark G, Borg-Karlson A-K, Stenlid J (2011) Chemical and transcriptional responses of Norway spruce genotypes with different susceptibility to *Heterobasidion* spp. infection. BMC Plant Biol 11(1):154. https://doi.org/10.1186/1471-2229-11-154

DeBolt S (2010) Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. Genome Biol Evol 2(0):441–453. https://doi.org/10.1093/gbe/evq033

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the *Photoperiod-B1* and *Vernalization-A1* genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS One 7(3):e33234. https://doi.org/10.1371/journal.pone.0033234

Duan J, Zhang JG, Deng HW, Wang YP (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PLoS One 8(3):e59128. https://doi.org/10.1371/journal.pone.0059128

Dube S, Qin J, Ramakrishnan R (2008) Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device. PLoS One 3(8):e2876. https://doi.org/10.1371/journal.pone.0002876

Erlandson KJ, Cotter CA, Charity JC, Martens C, Fischer ER, Ricklefs SM, Porcella SF, Moss B (2014) Duplication of the A17L locus of vaccinia virus provides an alternate route to rifampin resistance. J Virol 88(19):11576–11585. https://doi.org/10.1128/JVI.00618-14

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in

genome diversity. Genome Res 16(8):949–961. https://doi.org/10.1101/gr.3677206

Friedman AR, Baker BJ (2007) The evolution of resistance genes in multi-protein plant resistance systems. Curr Opin Genet Dev 17(6):493–499. https://doi.org/10.1016/j.gde.2007.08.014

Gaines TA, Zhang W, Wang D, Bukun B, Chisholm ST, Shaner DL, Nissen SJ, Patzoldt WL, Tranel PJ, Culpepper AS, Grey TL, Webster TM, Vencill WK, Sammons RD, Jiang J, Preston C, Leach JE, Westra P (2010) Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc Natl Acad Sci U S A 107(3): 1029–1034. https://doi.org/10.1073/pnas.0906649107

Ghosh S, Qu Z, Das PJ, Fang E, Juras R, Cothran EG, McDonell S, Kenney DG, Lear TL, Adelson DL, Chowdhary BP, Raudsepp T (2014) Copy number variation in the horse genome. PLoS Genet 10(10):e1004712. https://doi.org/10.1371/journal.pgen.1004712

Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. PathoGenetics 1(1):4. https://doi.org/10.1186/1755-8417-1-4

Hashemi J, Fotouhi O, Sulaiman L, Kjellman M, Höög A, Zedenius J, Larsson C (2013) Copy number alterations in small intestinal neuroendocrine tumors determined by array comparative genomic hybridisation. BMC Cancer 13(1):505. https://doi.org/10.1186/1471-2407-13-505

Hedman H, Zhu T, von Arnold S, Sohlberg JJ (2013) Analysis of the WUSCHEL-RELATED HOMEOBOX gene family in the conifer *Picea abies* reveals extensive conservation as well as dynamic patterns. BMC Plant Biol 13(1):89. https://doi.org/10.1186/1471-2229-13-89

Hu X, Worton RG (1992) Partial gene duplication as a cause of human disease. Hum Mutat 1(1):3–12. https://doi.org/10.1002/humu.1380010103

Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH (2004) Whole genome DNA copy number changes idetified by high density oligonucleotide arrays. Hum Genomics 1(4):287–299. https://doi.org/10.1186/1479-7364-1-4-287

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36(9):949–951. https://doi.org/10.1038/ng1416

Iovene M, Zhang T, Lou Q, Buell CR, Jiang J (2013) Copy number variation in potato – an asexually propagated autotetraploid species. Plant J 75(1):80–89. https://doi.org/10.1111/tpj.12200

Ju YS, Hong D, Kim S, Park SS, Kim S, Lee S, Park H, Kim JI, Seo J–S (2010) Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res 38(20):e190. https://doi.org/10.1093/nar/gkq730

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridisation for molecular cytogenetic analysis of solid tumors. Science 258(5083):818–821. https://doi.org/10.1126/science.1359641

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Exome Sequencing Project NHLBI, Quinlan AR, Nickerson DA, Eichler EE (2012) Copy number variation detection and genotyping from exome sequence data. Genome Res 22(8):1525–1532. https://doi.org/10.1101/gr.138115.112

Krutovsky KV, Elsik CG, Matvienko M, Kozik A, Neale DB (2006) Conserved ortholog sets in forest trees. Tree Genet Genomes 3(1): 61–70. https://doi.org/10.1007/s11295-006-0052-2

Kulka J, Tôkés AM, Kaposi-Novák P, Udvarhelyi N, Keller A, Schaff Z (2006) Detection of HER-2/neu gene amplification in breast carcinomas using quantitative real-time PCR—a comparison with immunohistochemical and FISH results. Pathol Oncol Res 12(4):197–204. doi: PAOR.2006.12.4.0197. https://doi.org/10.1007/BF02893412

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, Jia J, Kong X (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol 196(1):282–291. https://doi.org/10.1111/j.1469-8137.2012.04243.x

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS (2015) Early genome duplications in conifers and other seed plants. Sci Adv 1(10):e1501084. https://doi.org/10.1126/sciadv.1501084

Liu JJ, Ekramoddoullah AKM (2009) Identification and characterization of the WRKY transcription factor family in *Pinus monticola*. Genome 52(1):77–88. https://doi.org/10.1139/G08-106

Liu P, Carvalho CMB, Hastings PJ, Lupski JR (2012) Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev 22(3):211–220. https://doi.org/10.1016/j.gde.2012.02.012

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KCQ, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 13(10):2291–2305. https://doi.org/10.1101/gr.1349003

McHale LK, Haun WJ, WW X, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159(4):1295–1308. https://doi.org/10.1104/pp.112.194605

Mehta D, Iwamoto K, Ueda J, Bundo M, Adati N, Kojima T, Kato T (2014) Comprehensive survey of CNVs influencing gene expression in the human brain and its implications for pathophysiology. Neurosci Res 79:22–33. https://doi.org/10.1016/j.neures.2013.10.009

Ministry of Agriculture of the Republic of Latvia (2014) Latvian Forest Sector in Facts and Figures, NGO "Zaļās Mājas", Riga https://www.zm.gov.lv/public/ck/files/ZM/mezhi/buklets/Latvian_Forest_Sector_in_Facts_and_Figures2014.pdf

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 8(2):122–128. https://doi.org/10.1016/j.pbi.2004.12.001

Neiman M, Olson MS, Tiffin P (2009) Selective histories of poplar protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates. New Phytol 183(3):740–750. https://doi.org/10.1111/j.1469-8137.2009.02936.x

Nystedt B, Street NR, Watterbom A et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451):579–584. https://doi.org/10.1038/nature12211

Panchy N, Lehti-Shiu M, Shiu SH (2016) Evolution of gene duplication in plants. Plant Physiol 171(4):2294–2316. https://doi.org/10.1104/pp.16.00523

Pearce S, Saville R, Vaughan SP, Chandler PM, Wilhelm EP, Sparks CA, Al-Khaff N, Korolev A, Boulton MI, Phillips AL, Hedden P, Nicholson P, Thomas SG (2011) Molecular characterisation of Rht-1 dwarfing genes in hexaploid wheat. Plant Physiol 157(4): 1820–1831. https://doi.org/10.1104/pp.111.183657

Perry GH, Ben - Dor A, Tsalenko A, Sampas N, Rodriguez - Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82(3):685–695. https://doi.org/10.1016/j.ajhg.2007.12.010

Prunier J, Caron S, MacKay J (2017) CNVs into the wild: screening the genomes of conifer trees (*Picea spp.*) reveals fewer gene copy number variations in hybrids and links to adaptation. BMC Genomics 18(1):97. https://doi.org/10.1186/s12864-016-3458-8

Raff ML, Craigen WJ, Smith LT, Keene DR, Byers PH (2000) Partial COL1A2 gene duplication produces features of osteogenesis imperfecta and Ehlers-Danlos syndrome type VII. Hum Genet 106(1):19–28. https://doi.org/10.1007/s004390051004

Rajaraman J, Douchkov D, Lueck S, Hensel G, Nowara D, Pogoda M, Rutten T, Meitzel T, Hoefle C, Hueckelhoven R, Klinkenberg J, Trujillo M, Bauer E, Schmutzer T, Himmelbach A, Mascher M, Lazzari B, Stein N, Kumlehn J, Schweizer P (2017) The partial duplication of an E3-ligase gene in Triticeae species mediates resistance to powdery mildew fungi. BioRxiv. https://doi.org/10.1101/190728

Salojärvi J, Smolander OP, Nieminen K, Rajaraman S, Safronov O, Safdari P, Lamminmäki A, Immanen J, Lan T, Tanskanen J, Rastas P, Amiryousefi A, Jayaprakash B, Kammonen JI, Hagqvist R, Eswaran G, Ahonen VH, Serra JA, Asiegbu FO, de Dios Barajas-Lopez J, Blande D, Blokhina O, Blomster T, Broholm S, Brosché M, Cui F, Dardick C, Ehonen SE, Elomaa P, Escamez S, Fagerstedt KV, Fujii H, Gauthier A, Gollan PJ, Halimaa P, Heino PI, Himanen K, Hollender C, Kangasjärvi S, Kauppinen L, Kelleher CT, Kontunen-Soppela S, Koskinen JP, Kovalchuk A, Kärenlampi SO, Kärkönen AK, Lim KJ, Leppälä J, Macpherson L, Mikola J, Mouhu K, Mähönen AP, Niinemets Ü, Oksanen E, Overmyer K, Palva ET, Pazouki L, Pennanen V, Puhakainen T, Poczai P, Possen BJHM, Punkkinen M, Rahikainen MM, Rousi M, Ruonala R, van der Schoot C, Shapiguzov A, Sierla M, Sipilä TP, Sutela S, Teeri TH, Tervahauta AI, Vaattovaara A, Vahala J, Vetchinnikova L, Welling A, Wrzaczek M, Xu E, Paulin LG, Schulman AH, Lascoux M, Albert VA, Auvinen P, Helariutta Y, Kangasjärvi J (2017) Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet 49(6):904–912. https://doi.org/10.1038/ng.3862

Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 3(6):1101–1108. https://doi.org/10.1038/nprot.2008.73

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528. https://doi.org/10.1126/science.1098918

Šķipars V, Krivmane B, Ruņģis D (2011) Thaumatin-like protein gene copy number variation in scots pine (*Pinus sylvestris*). Environmental and. Exp Biol 9:75–81

Šķipars V, Šņepste I, Krivmane B, Veinberga I, Ruņģis D (2014) A method for isolation of high-quality total RNA from small amounts of woody tissue of Scots pine. Balt For 20(2):230–237

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5(11):e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61(1):437–455. https://doi.org/10.1146/annurev-med-100708-204735

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81. https://doi.org/10.1038/nature15394

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318(5855):1446–1449. https://doi.org/10.1126/science.1146853

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20(12):1689–1699. https://doi.org/10.1101/gr.109165.110

Toll-Riera M, Laurie S, Radó-Trilla N, Alba MM (2011) Partial gene duplication and the formation of novel genes. In: Friedberg F (ed) Gene duplication. InTech, Rijeka, pp 97–110. doi: https://doi.org/10.5772/21846. Available from: https://www.intechopen.com/books/gene-duplication/partial-gene-duplication-and-the-formation-of-novel-genes

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37(7):727–732. https://doi.org/10.1038/ng1562

Voronova A, Rungis D (2014) Development and characterisation of IRAP markers from expressed retrotransposon-like sequences in *Pinus sylvestris* L. Proceedings of the Latvian Academy of Sciences. Section B. Natural, Exact, and Applied Sciences 67(6):485–492. https://doi.org/10.2478/prolas-2013-0082

Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Paterson AH (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS One 6(12):e28150. https://doi.org/10.1371/journal.pone.0028150

Wang H, Nettleton D, Ying K (2014) Copy number variation detection using next generation sequencing read counts. BMC Bioinformatics 15(1):109. https://doi.org/10.1186/1471-2105-15-109

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP, Mohamadi H, Paulino D, Chiu R, Jackman SD, Robertson G, Yang C, Boyle B, Hoffmann M, Weigel D, Nelson DR, Ritland C, Isabel N, Jaquish B, Yanchuk A, Bousquet J, Jones SJM, MacKay J, Birol I, Bohlmann J (2015) Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J 83(2):189–212. https://doi.org/10.1111/tpj.12886

Wingen LU, Münster T, Faigl W, Deleu W, Sommer H, Saedler H, Theißen G (2012) Molecular genetic basis of pod corn (Tunicate maize). Proc Natl Acad Sci U S A 109(18):7715–7720. https://doi.org/10.1073/pnas.1111670109

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18(6):292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg SL (2017) An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 6(1):1–4. https://doi.org/10.1093/gigascience/giw016

Zmienko A, Samelak-Czajka A, Kozlowski P, Szymanska M, Figlerowicz M (2016) *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning *MSH2*, *AT3G18530* and *AT3G18535* genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. BMC Genomics 17(1):893. https://doi.org/10.1186/s12864-016-3221-1