A proposed naming convention for genes in Rosaceae species

Rosaceae Gene Name Standardization Subcommittee of RosEXEC and RosIGI: Sook Jung[1], Carole Bassett[2], Doug Bielenberg[3], Chun-Huai Cheng[1], Chris Dardick[2], Dorrie Main[1], Lee Meisel[4], Janet Slovin[5], Michela Troggio[6] and Robert J. Schaffer[7]

[1]Washington State University,Pullman, WA, USA , [2]USDA ARS Appalachian Fruit Research Station, Kearneysville, WV, USA, ARS, [3]Clemson University, Clemson SC, USA, [4]University of Chile, Santiago, Chile, [5]USDA, Beltsville, MD, USA, [6]Istituto Agrario San Michele all'Adige, San Michele all'Adige, Italy, [7]Plant & Food Research Ltd., Auckland, New Zealand

Introduction

The importance of a consistent hierarchical naming convention in science was pioneered by Linnaus, who was the father of bringing order into biological species. Through linking similarity in species he generated a naming convention that is still used today. Naming gives clarity as well as and a small amount of biological information to researchers, with genes in model species often being named after their mutant phenotype. The convention has been to generate a 2-3 letter acronym for each gene and when comparing across different species usually a 2 letter species abbreviation is used as a prefix. For example, the prefix for *Arabidopsis thaliana* genes would be At. The species prefix is not required when presenting results from a single species but it is necessary when genes from multiple species are being compared.

Naming rights are typically given to the "discoverer", and like all things in history, sometimes things are discovered more than once. In model species such as Arabidopsis and tomato, multiple names have been used in literature for a single gene. For example ETHYLENE INSENSITIVE 5 has also been published as AIN1 and XRN4. This has only been resolved by the genome naming convention of giving genes a chromosomal name and gene number (in this case AT1G54490) for the predicted genes from whole genome sequencing. The Rosaceae community is significantly smaller than the Arabidopsis community, and therefore we have the chance to set out naming conventions that do not fall into such traps.

Proposed Naming Convention

1. Species abbreviation

The comparison of genes across Rosaceae species gives valuable insights into the way that different species have evolved. Whole genome sequence is available for strawberry (Shulaev et al. 2010), apple (Velasco et al. 2010), peach (International Peach Genome Initiative 2013), and pear (Wu et al. 2013), with more to come. This allows cross species comparisons at the genome level and within a gene family. Considering the 3000 species within Rosaceae, however, it is clear that two-letter species abbreviations will not give clear distinction between species. It is likely that the majority of researchers will be focusing their gene based research on commercially cultivated varieties, however, even within these there are issues as both *Prunus persica* and *Pyrus pyrofolia* would become Pp. We therefore propose standard naming of major Rosaceae species using the following abbreviations (Table 1). Using a 3-letter prefix resulted in two conflicts: both *Prunus cerasus* and *Prunus cerasifera* would be Pce following our convention, so we recommend Pci for *Prunus cerasifera*. Likewise, both *Prunus mume* and *Prunus munsoniana* would be Pmu. We recommend Pmn for *Prunus munsoniana* (highlighted in the table). For taxonomy studies across non commercial species, abbreviations will be needed to be longer to distinguish the species. For these papers we suggest that researchers take a UNIPROT approach (http://www.uniprot.org/docs/speclist) using 5 letter abbreviations: 3 for the Genus name and 2 for the species with *Malus* x *domestica* becoming Maldo and *Prunus persica* becoming Prupe. This convention is still not sufficient for the complete list, and in these papers a clear nomenclature needs to be stated in the paper by the researcher. **However, we recommend that authorsdo not include a species prefix in the gene symbol when they submit the gene data to NCBI, GDR or any other databases, to minimize the creation of duplicated names due to the differences in the species prefix.** In GDR, the gene symbol with the recommended three-letter prefix will be stored along with the gene symbol without the prefix for genes from the major varieties.

Table 1. Proposed Species Abbreviation for major Rosaceae species.

Genus	Species	Prefix
Aronia	melanocarpa	Ame
Chaenomeles	japonica	Cja
Cydonia	oblonga	Cob
Eriobotrya	japonica	Eja
Fragaria	x ananassa	Fan
Fragaria	chiloensis	Fch
Fragaria	vesca	Fve
Fragaria	virginiana	Fvi
Malus	x domestica	Mdo
Mespilus	germanica	Mge
Malus	pumila	Mpu
Malus	sylvestris	Msy
Prunus	americana	Pam
Prunus	angustifolia	Pan
Prunus	armeniaca	Par
Prunus	avium	Pav
Pyrus	calleryana	Pca
Prunus	cerasifera	Pci
Prunus	cerasus	Pce
Pyrus	communis	Pco
Prunus	domestica	Pdo
Prunus	dulcis	Pdu
Prunus	hortulana	Pho
Prunus	mume	Pmu
Prunus	munsoniana	Pmn
Prunus	nigra	Pni
Prunus	persica	Ppe
Pyrus	pyrifolia	Ppy
Prunus	salicina	Psa
Prunus	serotina	Pse
Prunus	simonii	Psi
Prunus	spinosa	Psp
Pyrus	ussuriensis	Pus
Rubus	idaeus	Rid
Rubus	occidentalis	Roc

2. Gene symbol

For the actual name we encourage using a 'root' symbol for members of a gene family together with a hierarchical numbering system. It is ideal to design the symbol sothat it can be associated with biological function. The Arabidopsis community has set out namingguidelines that can be found at (http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp). When giving a gene a name we encourage people to check that the gene they want to name (and publish) does not already have a name assigned by searching the current literature and gene databases (GenBank and GDR). Co-naming with Arabidopsis may be difficult due to clade expansions between the two species. In spite of this we recommend using similar naming and numbering when possible.

No specific naming convention is proposed for homeologues, but we recommend that they be named sequentially within the gene families. We propose the following convention for naming splice variants and alleles.

- When the gene name does not contain a number eg. *PG*
  - *PG1*, *PG2*... (homeologues and homologues)
- When the gene has a number in the name eg *DHN3*
  - *DHN3.1, DHN3.2* (homeologues and homologues)
- *DHN3.1_1, DHN3.1_2* (splice variant)
- *DHN3.1a, DHN3.1b* (allele)

Facilitation of Gene Naming Standardization

Submission of gene data to GDR prior to publication

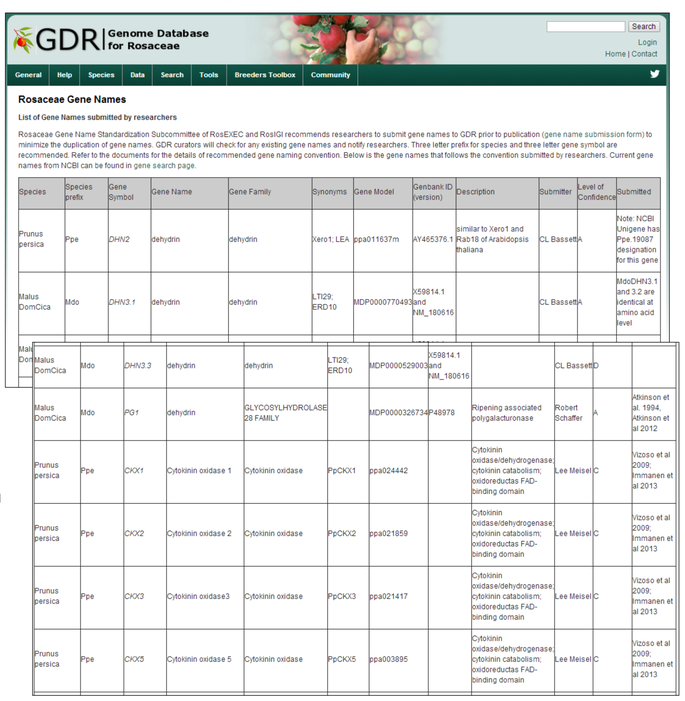
Researchers are encouraged to submit their gene data to GDR in addition to NCBI. While NCBI does not accept named genes that do not come from single molecule sequencing, the GDR database will. We also recommend researchers provide the corresponding gene models from whole genome sequencing, the gene family name, and the level of confidence based on the criteria we developed (see below**). When new gene models, gene sequences and/or splice variants are identified that are different from the whole genome data, we will recommend that users submit FASTA file in addition to the information in the table below.**

Species	Species prefix	Gene Symbol	Gene Name	Gene Family	Synonyms	Gene Model	Genbank ID	Description	Submitter	Level of confidence	Comments

Defining the 'level of confidence' for various sequencing methods and subsequent analysis.

A. Single molecule bi-directional cDNA sequence
B. Compilation DNA sequencing
C. Compilation of RNA-seq
D. Computational evidence

A GDR page for browsing expert-curated Rosaceae genes/gene families



Conclusions

Here we propose a standardised naming convention for Rosaceae species. We respectfully urge the Rosaceae community to follow these suggestions and by doing so, we will benefit from simplified literature reading, and less confusion when looking at genes. Please send feedback on these suggestions to any of the co-authors.