Potato Genome Sequencing Consortium Public Data Release

Sequence files and other related information for the Potato Genome Sequencing Consortium (PGSC). The PGSC has sequenced two potato species: the heterozygous diploid *S. tuberosum* Group Tuberosum cultivar, RH89-039-16 (RH), and the doubled monoploid *S. tuberosum* Group Phureja clone DM1-3 (DM)

- **Citations:**

For publication using the v4.04 pseudomolecules, please cite the following article:

Michael Alan Hardigan, Emily Crisovan, John P Hamiltion, Jeongwoon Kim, Parker Laimbeer, Courtney P Leisner, Norma C Manrique-Carpintero, Linsey Newton, Gina M Pham, Brieanne Vaillancourt, Xueming Yang, Zixian Zeng, David Douches, Jiming Jiang, Richard E Veilleux, and C. Robin Buell. 2016, Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum.* Plant Cell, doi:10.1105/tpc.15.00538
View the article here.

For publication using the PGSC v4.03 pseudomolecules, please cite the following two articles:

Potato Genome Sequencing Consortium 2011, Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.
View the article here.

Sharma, S. K., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., Carboni, M. F., D'Ambrosio, J. M., de la Cruz, G., Di Genova, A., Douches, D. S., Eguiluz, M., Guo, X., Guzman, F., Hackett, C. A., Hamilton, J. P., Li, G., Li, Y., Lozano, R., Maass, A., Marshall, D., Martinez, D., McLean, K., Mejía, N., Milne, L., Munive, S., Nagy, I., Ponce, O., Ramirez, M., Simon, R., Thomson, S. J., Torres, Y., Waugh, R., Zhang, Z., Huang, S., Visser, R. G. F., Bachem, C. W. B., Sagredo, B., Feingold, S. E., Orjeda, G., Veilleux, R. E., Bonierbale, M., Jacobs, J. M. E., Milbourne, D., Martin, D. M. A. & Bryan, G. J. 2013, Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps. G3: Genes|Genomes|Genetics 3: 2031-2047.
View the article here.

- **Updates:**
- February 1, 2016 - The DM v4.04 pseudomolecules are available to download and search on the SpudDB BLAST server. More information about v4.04 can be found in the Genome Assemblies section below.
- December 13, 2013 - DM and RH RNA-Seq FPKM summary files regenerated with the corrected PGSC v4.03 pseudomolecule annotation
- December 5,2013 - The PGSC DM v4.03 pseudomolecule GFF3 files have been updated to correct an error. An error occurred when converting positions of annotations on the PGSC DM v3 superscaffolds to the PGSC DM v4.03 pseudomolecules if a superscaffold was split. This error affected 1628 genes, which are listed in this file. The error only affected the location of annotations and not the sequence. The FASTA and AGP files for the PGSC DM v4.03 pseudomolecules were not affected.
- November 1, 2013 - A paper describing the construction of the PGSC v4.03 pseudomolecules for *S. tuberosum* Group Phureja DM1-3 has been published in the journal *G3: Genes, Genomes, Genetics*.
- September 4, 2013 - The PGSC v4.03 pseudomolecule FASTA sequence, AGP, and GFF3 annotation is now available
- July 9, 2012 - The PGSC v2.1.10 pseudomolecules (based on version 3 of the DM genome assembly) were updated to v2.1.11 pseudomolecules. This new version is the same as the *S. tuberosum* Group Phureja DM1-3 Version 2.1.10 AGP Pseudomolecule Sequences (available below) except the gaps greater than 50 kbp have been changed to 50 kbp
- Dec 15, 2011 - The transcript and representative transcript files have updated due to the original files containingsome corrupted sequences.

Genome Assemblies (FASTA Format)

- The Buell lab at Michigan State have created a new pseudomolecule (chrUn) created from assembled DM reads that did not map to v4.03 and released it with the v4.03 chr00-chr12 pseudomolecules as v4.04. The pseudomolecules chr00-chr12 remain the same as v4.03. The v4.04 FASTA file can be downloaded below or searched on the SpudDB BLAST server. More details about the construction of chrUn can be found in the paper by Hardigan *et al.* (2016).

DM_v4.04_pseudomolecules.fasta.zip -
S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, Version 4.04 Pseudomolecule Sequence

- PGSC_DM_v4.03_pseudomolecules.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, PGSC Version 4.03 Pseudomolecule Sequence

  PGSC_DM_v4.03_unanchored_regions_chr00.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, PGSC Version 4.03 Chr00 Sequence (Unanchored Sequences)

  PGSC_DM_v4.03_pseudomolecules.agp.zip -
  S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, PGSC Version 4.03 Pseudomolecule AGP File

  PGSC_DM_v4.03_unanchored_regions_chr00.agp.zip -
  S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, PGSC Version 4.03 Chr00 AGP File (Unanchored Sequences)

- PGSC_DM_v3_2.1.11_pseudomolecules.zip -
  S. tuberosum Group Phureja DM1-3 Assembly Version 3 DM, PGSC Version 2.1.11 Pseudomolecule Sequences

  The PGSC v2.1.10 pseudomolecules (based on version 3 of the DM genome assembly) were updated to v2.1.11 pseudomolecules. This version is the same as the S. tuberosum Group Phureja DM1-3 Version 2.1.10 AGP Pseudomolecule Sequences (available below) except the gaps greater than 50 kbp have been changed to 50 kbp
- PGSC_DM_v3_2.1.10_pseudomolecules.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 DM, Version 2.1.10 AGP Pseudomolecule Sequences
- PGSC_DM_v3_2.1.10_superscaffolds_unanchored_gtr_2.5k.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 DM, Version 2.1.10 AGP Unanchored Superscaffold Sequences (>2.5kbp)
- PGSC_DM_v3_2.1.10_pseudomolecule_AGP.xlsx -
  S. tuberosum Group Phureja DM1-3 Version 3 DM Pseudomolecule AGP data (v2.1.10) - Excel Format
- PGSC_DM_v3_superscaffolds.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 DM superscaffold sequences
- PGSC_DM_v3_scaffolds.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 DM scaffold sequences
- S_tuberosum_Group_Phureja_chloroplast_DM1-3-516-R44.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 chloroplast sequences
- S_tuberosum_Group_Phureja_mitochondrion_DM1-3-516-R44.fasta.zip -
  S. tuberosum Group Phureja DM1-3 Version 3 mitochondrion sequences
- S_tuberosum_Group_Tuberosum_chloroplast_RH89-039-16.fasta.zip -
  S. tuberosum Group Tuberosum RH89-039-16 chloroplast sequences
- S_tuberosum_Group_Tuberosum_mitochondrion_RH89-039-16.fasta.zip -
  S. tuberosum Group Tuberosum RH89-039-16 mitochondrion sequences

S. tuberosum Group Phureja DM1-3 Genome Annotation v3.4 mapped to the pseudomolecule sequences

PGSC_DM_V403_genes.gff.zip -
Gene annotation for the v4.03 Pseudomolecules in GFF3 format

PGSC_DM_V403_representative_genes.gff.zip -
Representative gene annotation for the v4.03 Pseudomolecules in GFF3 format - Only the transcript that produces the longest peptide sequence among all the alternative isoforms of a gene is included.

- PGSC_DM_v3_2.1.11_pseudomolecule_annotation.gff.zip -
  Gene annotation for v2.1.11 Pseudomolecules in GFF3 format
- PGSC_DM_v3_2.1.10_pseudomolecule_annotation.gff.zip -
  Gene annotation for v2.1.10 Pseudomolecules in GFF3 format

S. tuberosum Group Phureja DM1-3 Genome Annotation v3.4 (based on v3 superscaffolds)

- PGSC_DM_v3.4_gene.fasta.zip -
  Nucleotide sequences of all genes.
- PGSC_DM_v3.4_cds.fasta.zip -
  Nucleotide sequences of all gene coding sequences (coding sequence only, i.e. no introns and no UTRs).
- PGSC_DM_v3.4_transcript-update.fasta.zip -
  Nucleotide sequences of all transcript sequences (UTRs and exons).
- PGSC_DM_v3.4_pep.fasta.zip -
  Amino acid sequences corresponding to all gene coding sequences.
- PGSC_DM_v3.4_gene.gff.zip -
  Gene annotation in GFF3 format
- PGSC_DM_v3.4_cds_nonredundant.fasta.zip -
  Alternative isoforms sometimes share the same coding sequence (CDS) which only appears once in this file.
- PGSC_DM_v3.4_pep_nonredundant.fasta.zip -
  Amino acid sequences corresponding to nonredundant CDS file above.

- PGSC_DM_v3.4_gene_nonredundant.gff.zip -
  Same as PGSC_DM_v3.4_gene.gff with additional flaggings for a) identical peptides originating from multiple genes b) identical peptides originating from alternative isoforms from the same gene.
- PGSC_DM_v3.4_transcript-update_representative.fasta.zip -
  The transcript that produces the longest peptide sequence among all the alternative isoforms of a gene is selected as the representative transcript.
- PGSC_DM_v3.4_cds_representative.fasta.zip -
  Coding sequences of the representative transcripts.
- PGSC_DM_v3.4_pep_representative.fasta.zip -
  Amio acid sequences corrsponding to the representative CDS file above
- PGSC_DM_v3.4_gene_func.txt.zip -
  Putative function of all genes. The putative function of the representative peptide is used if alternative isoforms exist.
- PGSC_DM_v3.4_g2t2c2p2func.txt -
  Linking file between gene ID, transcript ID, CDS ID, peptide ID, and putative function as determined by the representative peptide of the gene if alternative isoforms exist. This file includes all the transcripts in PGSC_DM_v3.4_gene.gff
- PGSC_DM_v3.4_g2t2c2p2func_nonredundant.txt.zip -
  Linking file between gene ID, transcript ID, CDS ID, peptide ID, and putative function as determined by the representative peptide of the gene if alternative isoforms exist. This file includes only the nonredundant transcripts.
- PGSC_DM_v3.4_representative_model.gtf.zip -
  GTF for version 3.4 representative models

Miscellaneous annotation based on the PGSC Version 4.03 Pseudomolecule

- PGSC_DM_V403_DArT.gff.zip -
  Unambiguously mapped Potato Diversity Arrays Technology (DArT) marker sequences on the PGSC Version 4.03 Pseudomolecules - GFF3 format
- PGSC_DM_V403_new_opa.gff.zip -
  Dundee-derived SNP marker positions on the PGSC Version 4.03 Pseudomolecules used for the Dundee oligo-nucleotide pooled assay (OPA) assay- GFF3 format
- PGSC_DM_V403_POPA_MICRO.gff.zip -
  Best BLASTN hits of oligo-nucleotide pooled assay (OPA) markers on the PGSC Version 4.03 Pseudomolecules. The file contains the marker sequences. - GFF3 format
- PGSC_DM_V403_SSR.gff.zip -
  Location of mapped simple sequence repeats (SSRs) markers on the PGSC Version 4.03 Pseudomolecules - GFF3 format
- PGSC_DM_V403_dmap_plotted_markers.gff.zip -
  Potato marker sequences plotted by DMAP on the PGSC Version 4.03 Pseudomolecules - GFF3 format
- PGSC_DM_V3_superscaffolds_miRNA.gff - Potato miRNAs (stem-loop precursor and mature) from miRBase v21 mapped to the PGSC v3 superscaffolds - GFF3 format
- PGSC_DM_V403_miRNA.gff - Potato miRNAs (stem-loop precursor and mature) from miRBase v21 mapped to the PGSC Version 4.03 Pseudomolecules - GFF3 format
- potato_dm_v4.03.putative.ssr.gff3.zip -
  Putative SSRs on the PGSC Version 4.03 Pseudomolecules - GFF3 format
- potato_dm_v4.03.repeatmasker.gff3.zip -
  RepeatMasker annotated repeats on the PGSC Version 4.03 Pseudomolecules - GFF3 format
- potato_69011SNPs_potato_dm_v4.03.gff3.zip -
  SolCAP Infinium High Confidence SNPs (http://solcap.msu.edu/potato_infinium.shtml) identified from Atlantic, Premier Russet, and Snowden RNA-seq and aligned to the PGSC Version 4.03 Pseudomolecules - GFF3 format
- potato_8303SNPs_potato_dm_v4.03.gff3.zip -
  SolCAP 8303 Array Infinium SNPs (http://solcap.msu.edu/potato_infinium.shtml) aligned to the PGSC Version 4.03 Pseudomolecules - GFF3 format
- RH_SNPs_vs_potato_dm_v4.03.gff3.zip -
  SNPs identified from aligning RH illumina reads to the PGSC v4.03 Pseudomolecules and calling SNPs with SAMTools variant calling pipeline - GFF3 format

RNA-Seq Gene Expression Data

- DM_RH_RNA-Seq_FPKM_expression_matrix_for_DM_v4.03_13dec2013_desc.xlsx -
  FPKM values of all the representative transcripts across 40DM and 16 RH libraries. - Excel File
- DM_RH_RNA-Seq_FPKM_expression_matrix_for_DM_v4.03_13dec2013_desc.txt.zip -
  FPKM values of all the representative transcripts across 40 DM libraries. - Tab Delimited File

Information about the RNA-Seq Gene Expression Data

The format of the files:
1st column: gene ID
2nd column: library 1
3rd column: library 2
...
last column: functional annotation of the gene

The reads were mapped to *S. tuberosum* Group Phureja DM1-3 superscaffolds using Tophat (v1.4.1) [which made use of Bowtie (v1.0.0)] The FPKM values were calculated by Cufflinks (v1.3.0) using v3.4 representative model set only.

Tophat was run with "-i 10 -I 15000" parameters, which set a minimum intron size of 10bp (-i 10), and a maximum intron size of 15,000bp (-I 15000). These values are the minimum and maximum intron feature lengths present in the v3.4 GFF. Paired-end libraries were aligned in single end mode.

Cufflinks was run with the same maximum intron size of 15,000bp (-I 15000)

Functional annotation was based on best BLASTX hits using the CDS sequences against UniRef100. The text was assigned using a first informative best-hit strategy, which considers best BLASTX hits where E <= 1e-5, but excludes hits with non-informative functional text (eg: "Whole genome shotgun sequence of line..."). The text is also programmatically cleaned to remove some misleading and low-information strings. For gene-level annotation, the transcript-level functional text was concatenated, so there will be some redundancy due to variations in the annotation string assigned to the different isoforms.

Putative Orthologous Groups (OrthoMCL)

- 12_plants_all_orthomcl_parsed.txt.zip -
  The predicted proteomes (representative peptides only) of 12 plant species were used for identification of putative orthologous groups using OrthoMCL with deafult parameters (Li et al., 2003). The plant species included are: *Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Solanum tuberosum, Sorghum bicolor, Vitis vinifera* and *Zea mays*.

This tab-delimited file has the following columns:
Cluster_ID,
Number_of_peptides_in_this_cluster
Number_of_species_in_this_cluster
Species (separated by space)
Peptides (separated by space)

BAC, BAC End, and Fosmid End Sequences

- Solanum_tuberosum.RH.bacs.zip -
  S. tuberosum Group Tuberosum RH89-039-16 BACs (RHPOTKEY library)
- Solanum_tuberosum.RH.bac_ends.zip -
  S. tuberosum Group Tuberosum RH89-039-16 BAC ends (RHPOTKEY library)
- Solanum_phureja.DM.bac_ends.zip -
  S. tuberosum Group Phureja DM1-3 BAC ends
- Solanum_phureja.DM.fosmid_ends.zip -
  S. tuberosum Group Phureja DM1-3 fosmid ends

Potato Diversity Arrays Technology (DArT) markers:

- Potato_DArT_sequences.tar.xz

The potato DArT array contains 7,680 probes obtained using genomic representations from a potato diversity panel also including selected probes from tomato (234) and Capsicum (54). The DArT probes were sequenced using financial support from The James Hutton Institute, UK under their Potato Genome Sequencing Grant* and are made available by Diversity Arrays Technology Pty Ltd, Yarralumla ACT 2600, Australia. This work is part of the Potato Mapping Group, a subgroup of the Potato Genome Sequencing Consortium (PGSC).

**Scottish Government Rural and Environmental Science and Analytical Services Division (RESAS), Department for Environment, Food and Rural Affairs (DEFRA), Agriculture and Horticulture Development Board (AHDB) - Potato Council.*

This work is supported by grants from the National Science Foundation (IOS- 2140176), U.S. Department of Agriculture (2019-51181-30021), and funds from the Georgia Research Alliance, Georgia Seed Development, and University of Georgia.