# Downstream Analysis of SNPs from the SolCAP Tomato Infinium Array (II)

Sung-Chur Sim

The Ohio State University, OARDC

2011 SolCAP Workshop

Mathilde Causse

Martin Ganal

# Overview

- SolCAP tomato SNPs
- How to select subsets of SNPs for mapping populations
  - Search for SNPs using parents
  - Obtain physical map positions for SNPs
  - Filter redundant SNPs based on map information
- Minor allele frequency (MAF) and linkage disequilibrium (LD) analysis
- Summary

# Learning Objectives

At the end of this presentation, you will be able to:

- Retrieve the SolCAP Cluster File
- Identify polymorphic SNPs in a specific mapping population
- Find the physical map locations of SNPs via BLAST
- Select non-redundant and informative SNPs

# Sample data

- Two sample data files for demonstration in the workshop URL (http://www.extension.org/pages/61007)
  - PMserach_SolCAP.xls
  - FASTAmultiquery_SolCAP.txt

- Download these files now if you want to follow “hands on” examples

- Open your browser to SGN (http://solgenomics.net/)

# SolCAP Tomato SNPs

- Illumina Infinium SNP chip assay
  - 8,784 genome-wide SNPs with 10,000 probes
  - 7,720 SNPs passed production QC
  - 3-day procedure per run (24–192 samples) using iScan
  - Genotyping facility at MSU and UC-Davis

- SolCAP germplasm panel (n=489) *
  - Processing –141 accessions
  - Fresh market – 122 accessions
  - Vintage – 88 accessions
  - Wild – 103 accessions
  - Others (hybrids, $F_2$ etc.) – 35 accessions

* Number of accessions that we genotyped as of 09/27/2011

# SNP Calls in the Tomato Panel

- 36 accessions were duplicated for QC
  - Perfect match – 34 (94.0%)
  - Inconsistent calls – 2 (6.0%)
    - ✓ Principle Borghese and Ailsa Craig had high level of Heterozygous calls that were inconsistent between DNA sources
- SNP calls from 7,720 SNPs in the tomato panel
  - <10% missing – 7,535 (97.6%)
  - 10-19% missing – 92 (1.2%)
  - ≥ 20% missing – 57 (0.7%)
  - No calls – 36 (0.5%)
- SNP data for 141 processing accessions is available on the workshop URL (http://www.extension.org/pages/61007): ProcessingData_SolCAP.xlsx

# Cluster File for SNP Calling

- SolCAP developed a Cluster File (version 1) for SNP calling of the tomato Infinium array
- The Cluster File is based on the SolCAP tomato panel consisting of mostly inbreds and a few hybrids
- The Cluster File is now available on the workshop URL (http://www.extension.org/pages/61007):
  - SolCAP_ClusterFile_v1.egt
  - **Note: DATA ACCESS AGREEMENT**

# Applications of Tomato Infinium Data

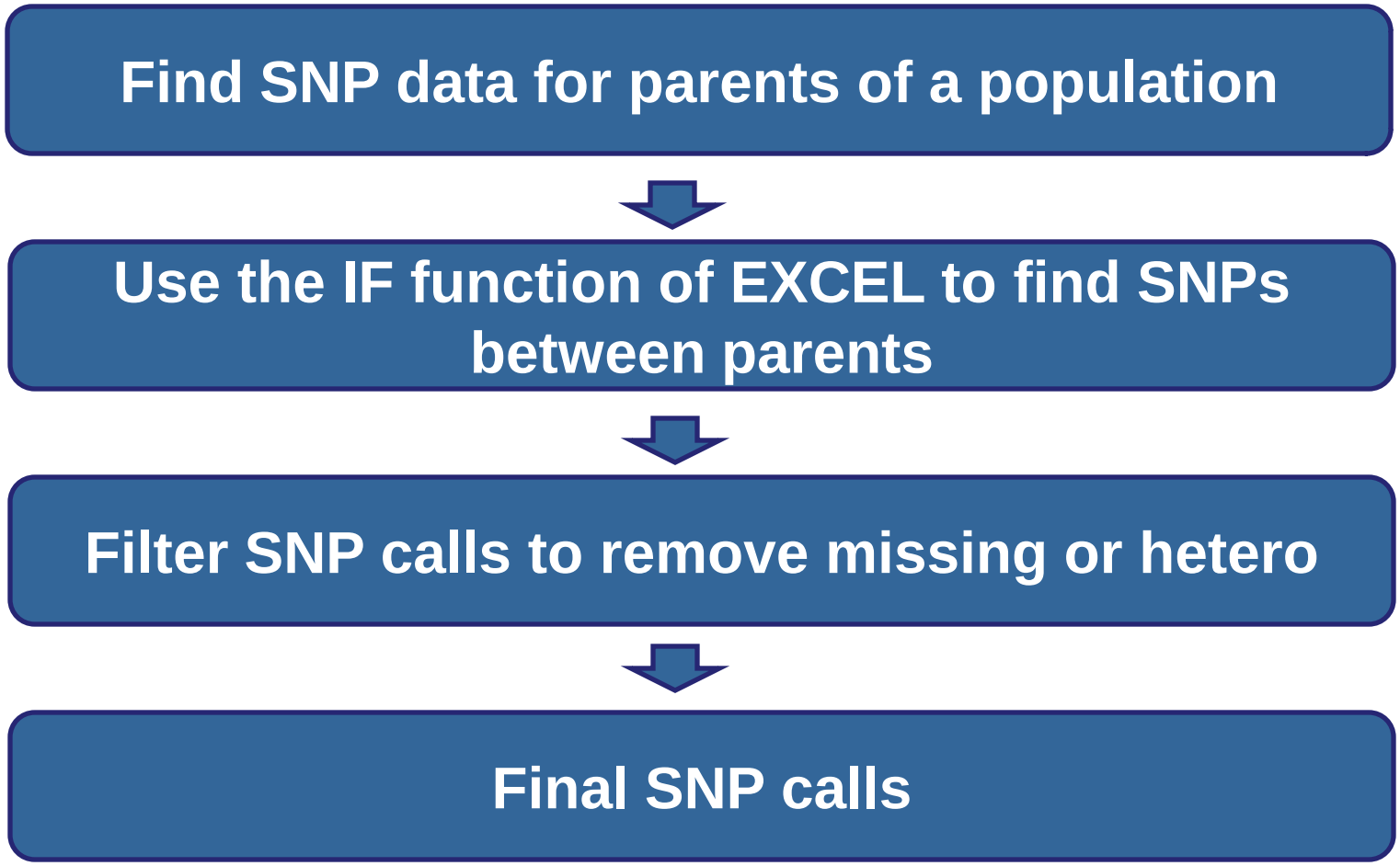
- Excellent survey tool to identify polymorphic SNPs in cultivated germplasm and wild germplasm

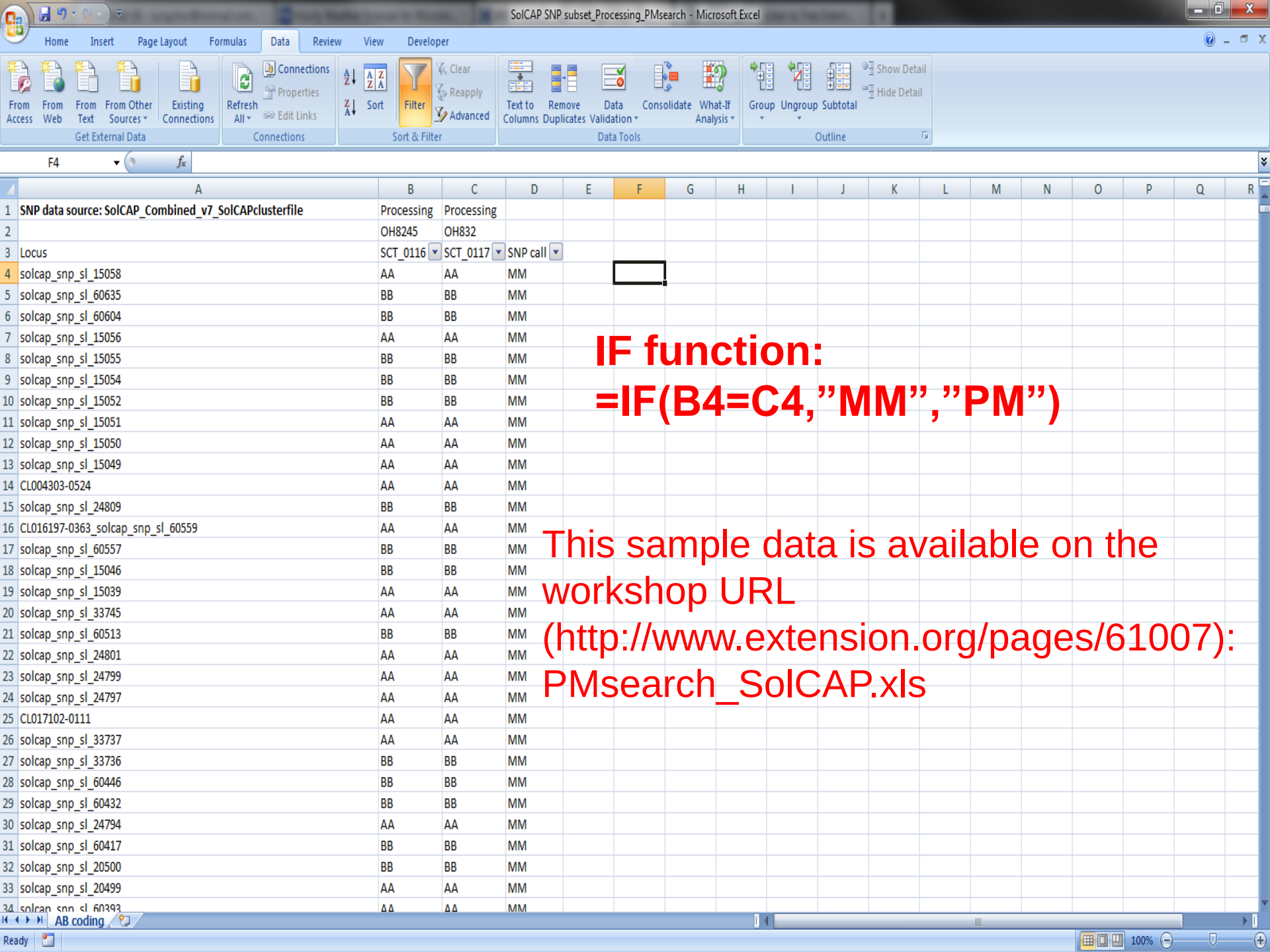
- Genomic resources to ask biological questions (e.g. How is variation distributed within and between market classes?) via the following analysis
  - Allele frequency analysis
  - LD analysis
  - Population structure analysis
  - Association analysis

# Data Handling

- Data manipulation using spreadsheet
  - Excel functions (IF, COUNTIF, etc)
- The use of R is an ideal option for a large data
- Resources for data manipulation in R
  - Introduction to R Statistical Software Webinar presented by Heather Merk (http://www.extension.org/pages/60427)
  - Kim, D.Y. R basics [Online]. Illinois State University (http://math.illinoisstate.edu/dhkim/rstuff/rtutor.html )
  - R package ‘gdata’ (http://cran.r-project.org/web/packages/gdata/ )

# How to Identify Polymorphic SNPs for a Specific Population






**IF function:**
**=IF(B4=C4,"MM","PM")**

This sample data is available on the workshop URL (http://www.extension.org/pages/61007): PMsearch_SolCAP.xls

# How to Obtain Physical Map Positions of the Tomato SNPs

Download the file ‘Tomato Infinium SNP Annotation’ at the SolCAP website (http:// solcap.msu.edu/tomato_genotype_data.shtml )

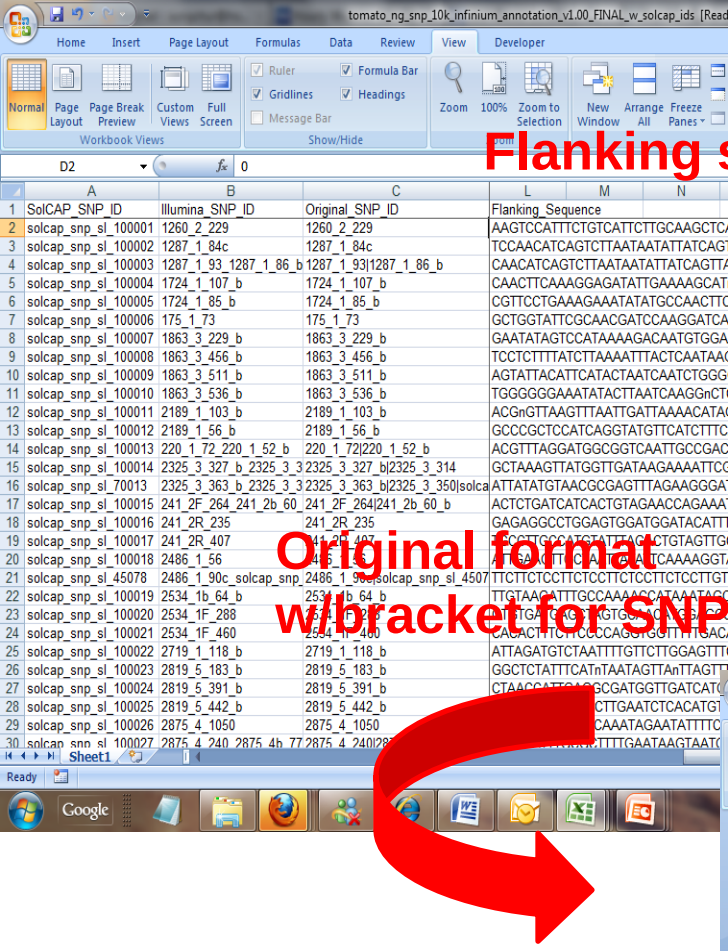
From the file, copy and paste flanking sequences in a new spreadsheet in order to make a multi-fasta query file for BLAST

Run BLAST against the tomato genome sequences at SGN (http://solgenomics.net )

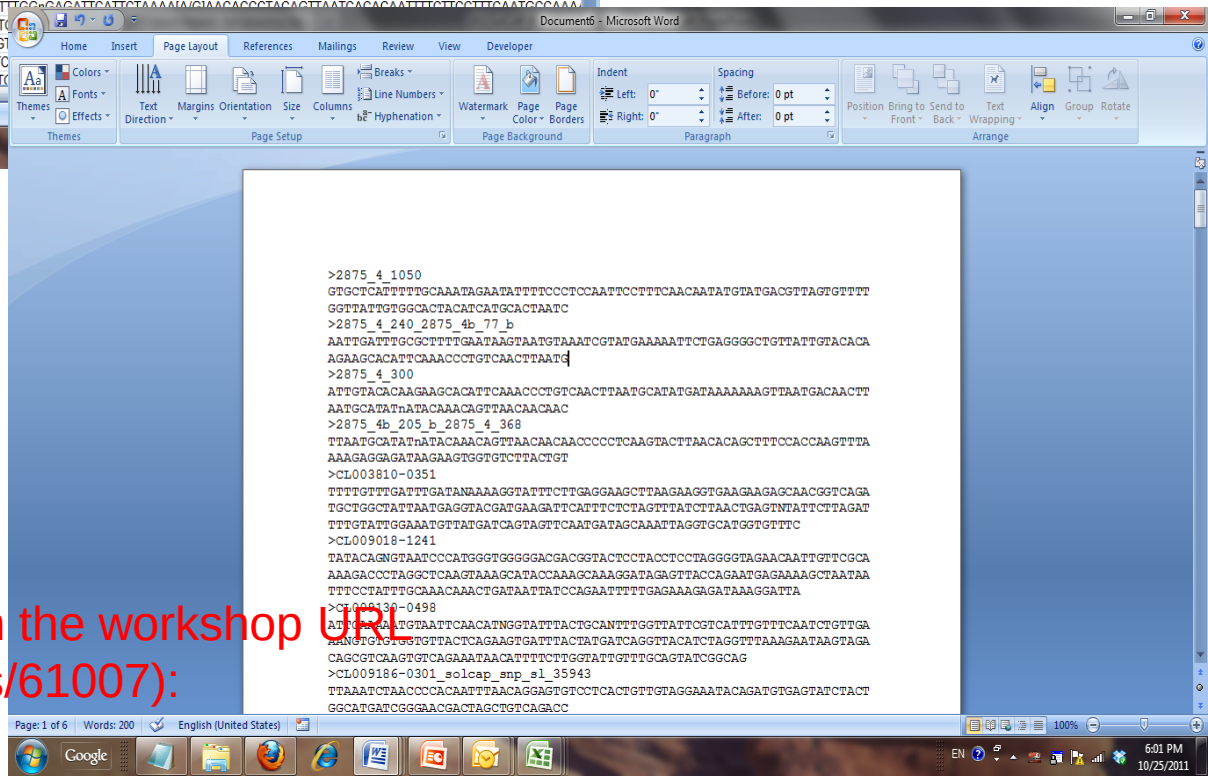
# Flanking sequences of SNPs

‘Tomato Infinium SNP Annotation’ file

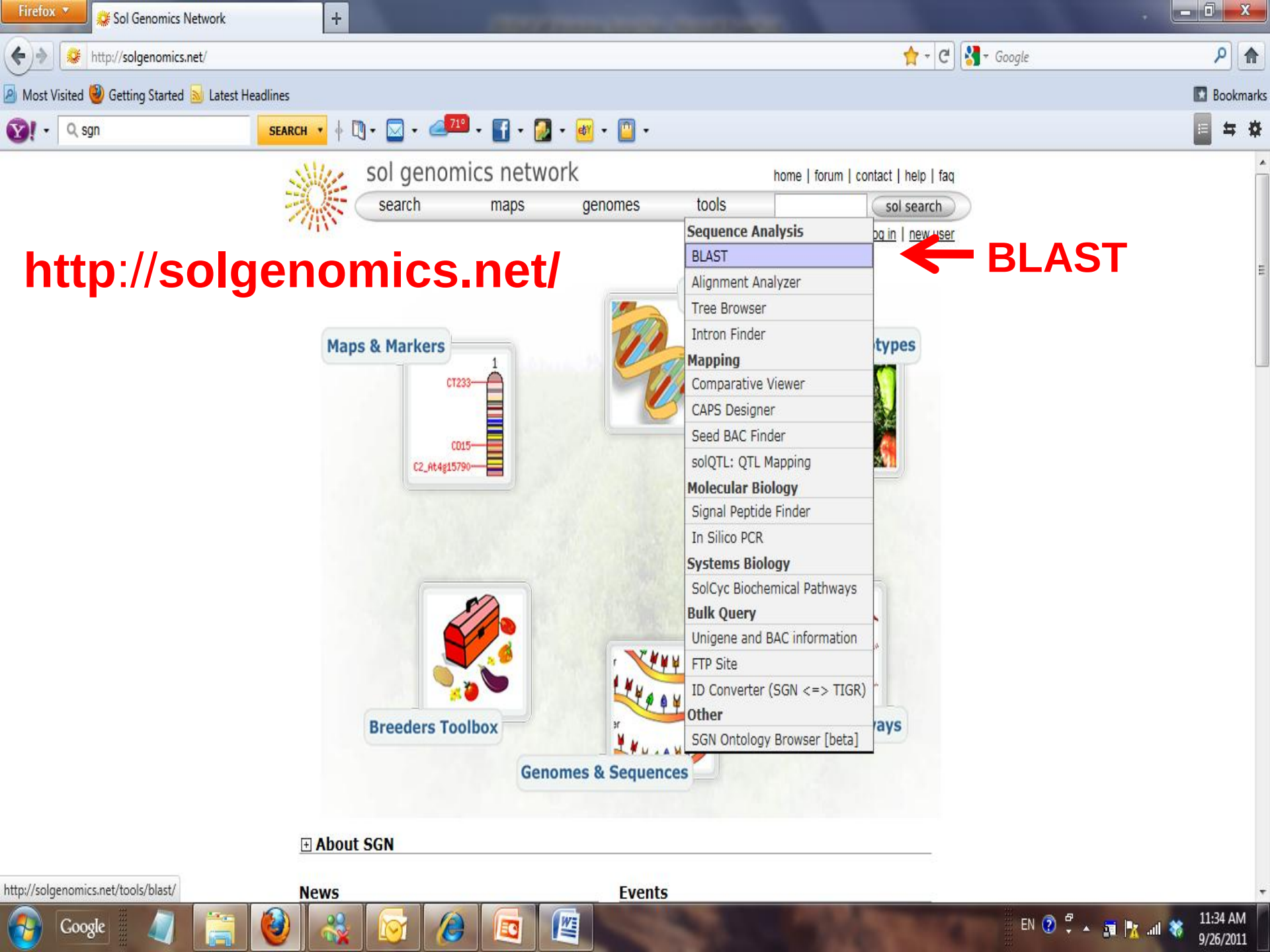
Original format w/bracket for SNP

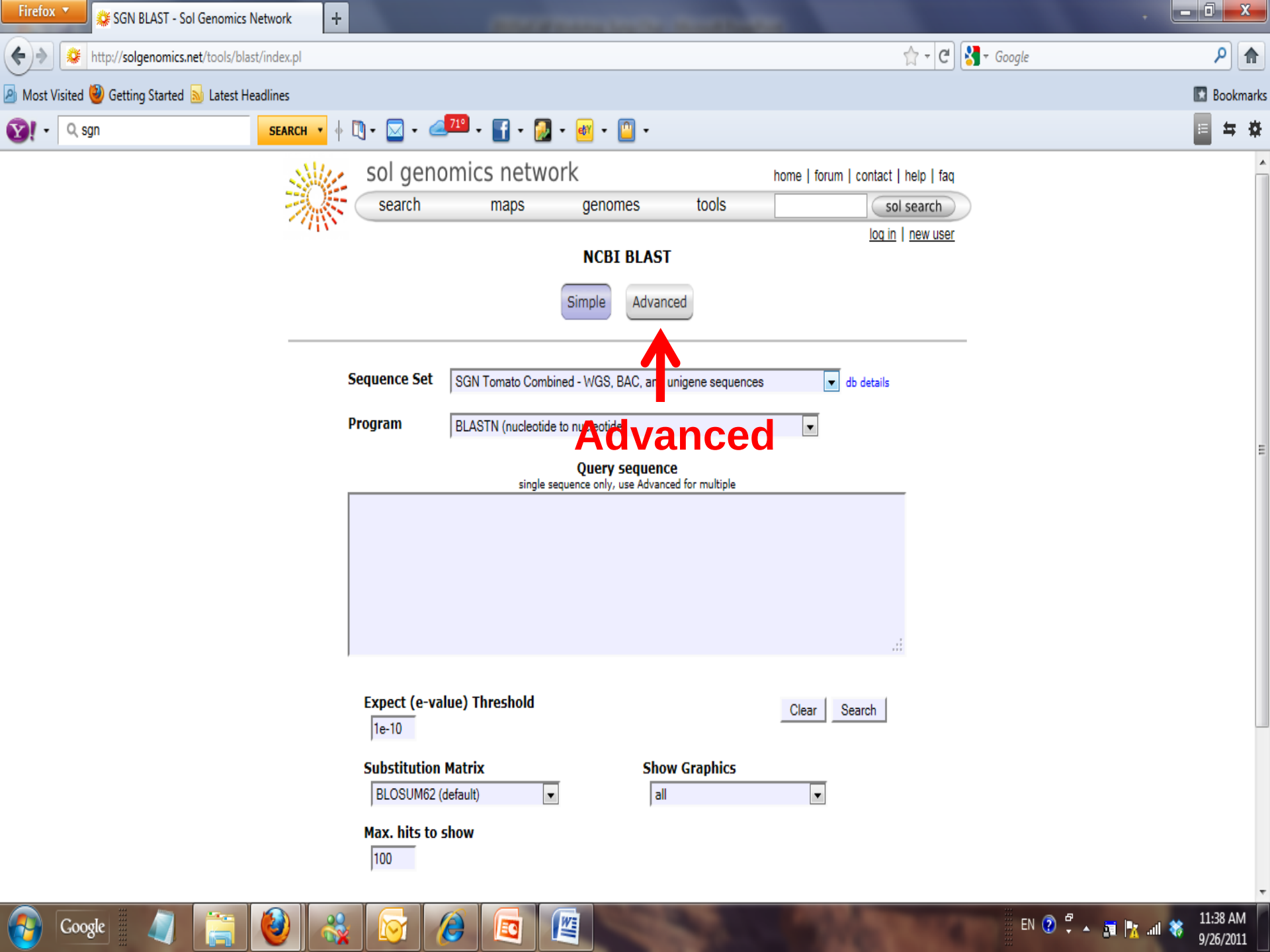


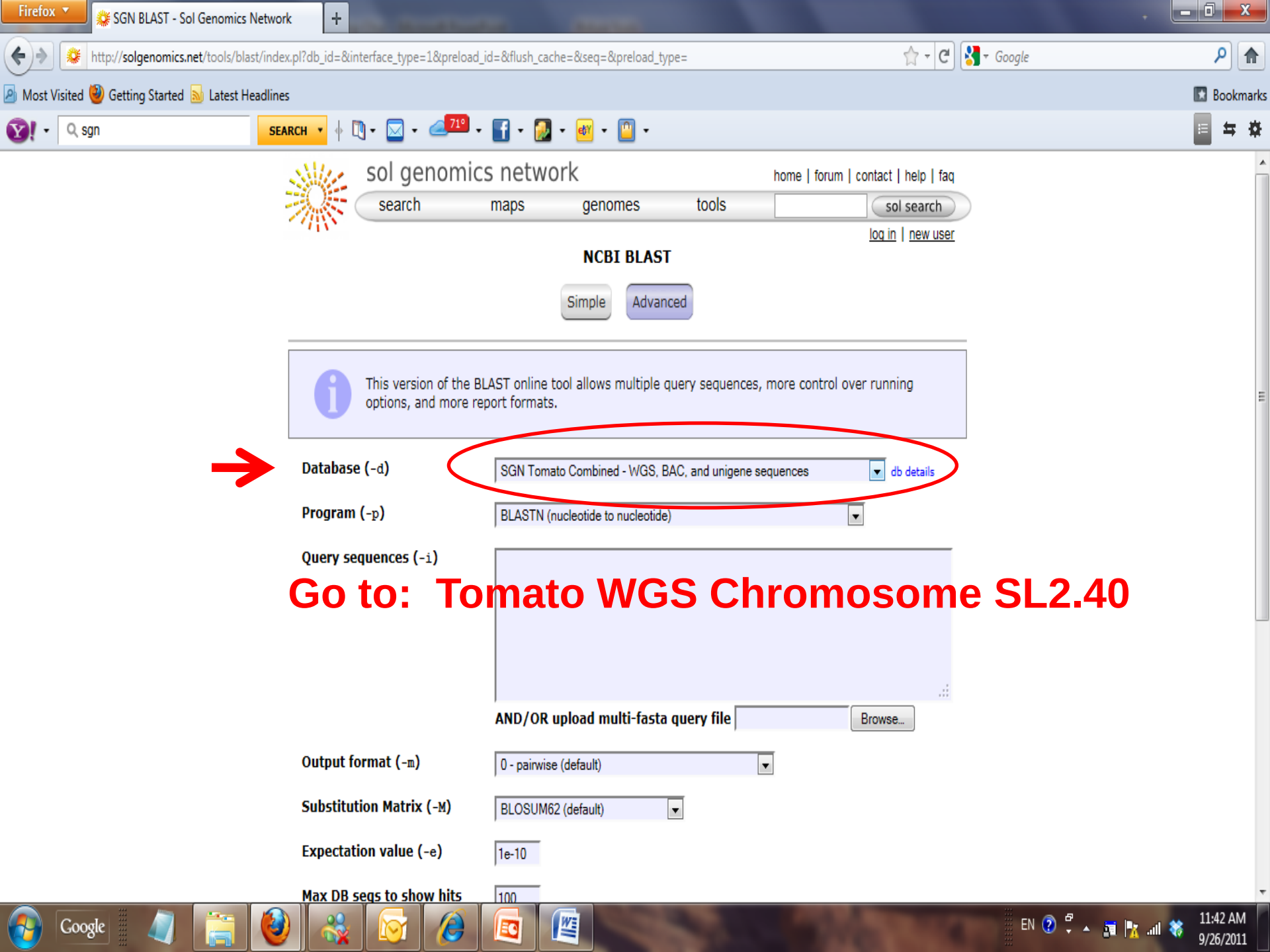

FASTA format

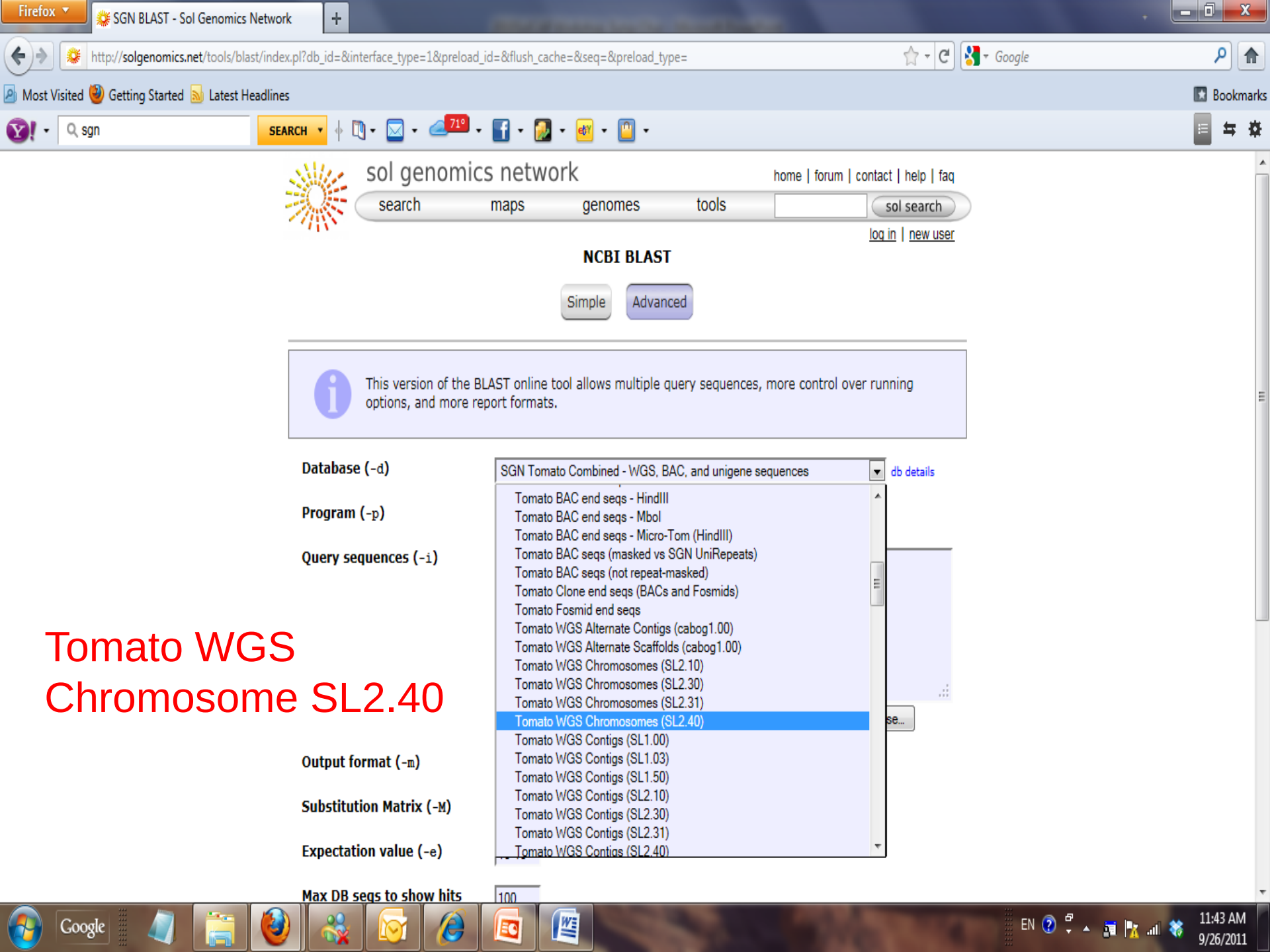


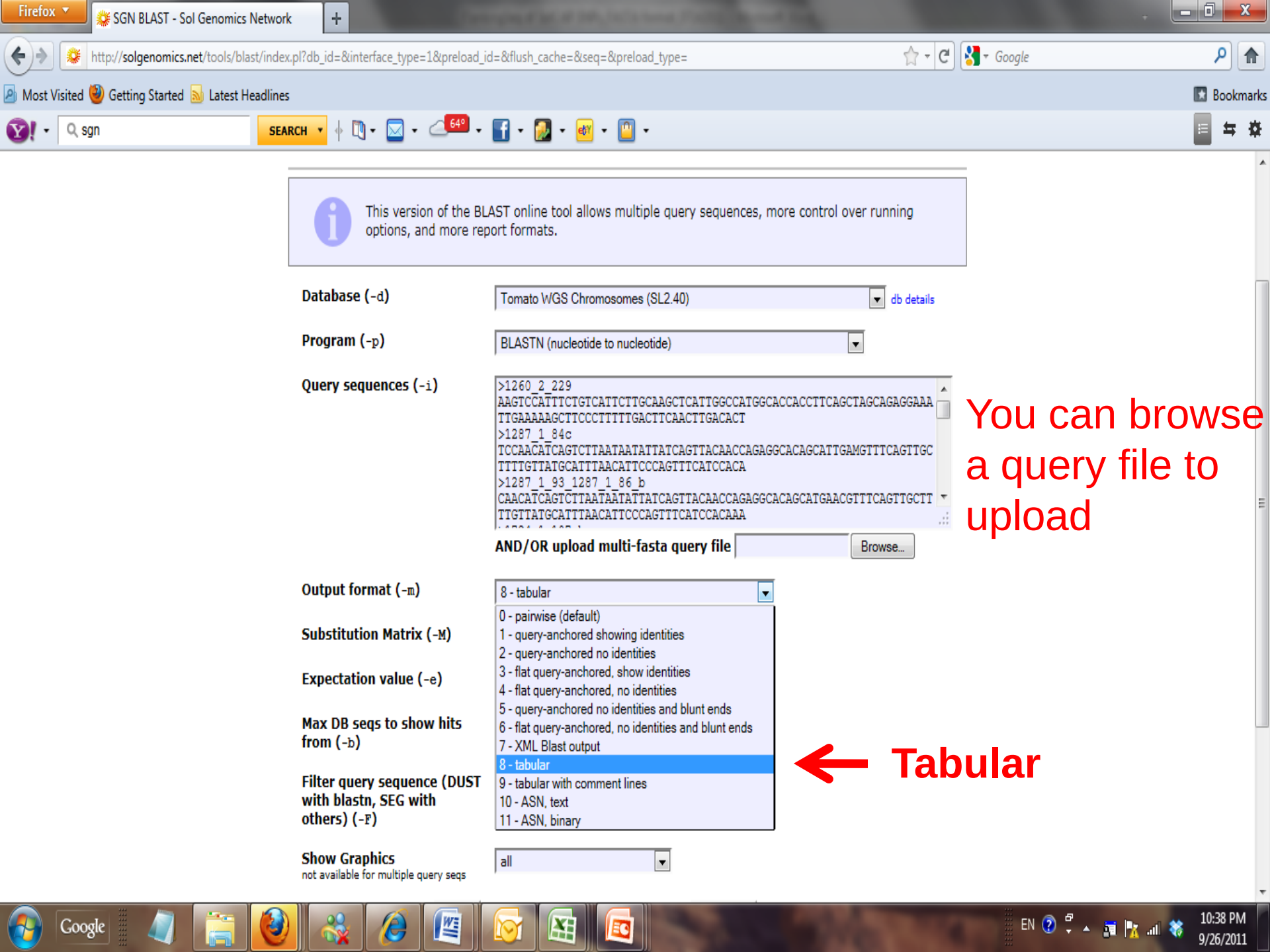

This sample data is available on the workshop URL (http://www.extension.org/pages/61007): FASTAmultiquery_SolCAP.txt

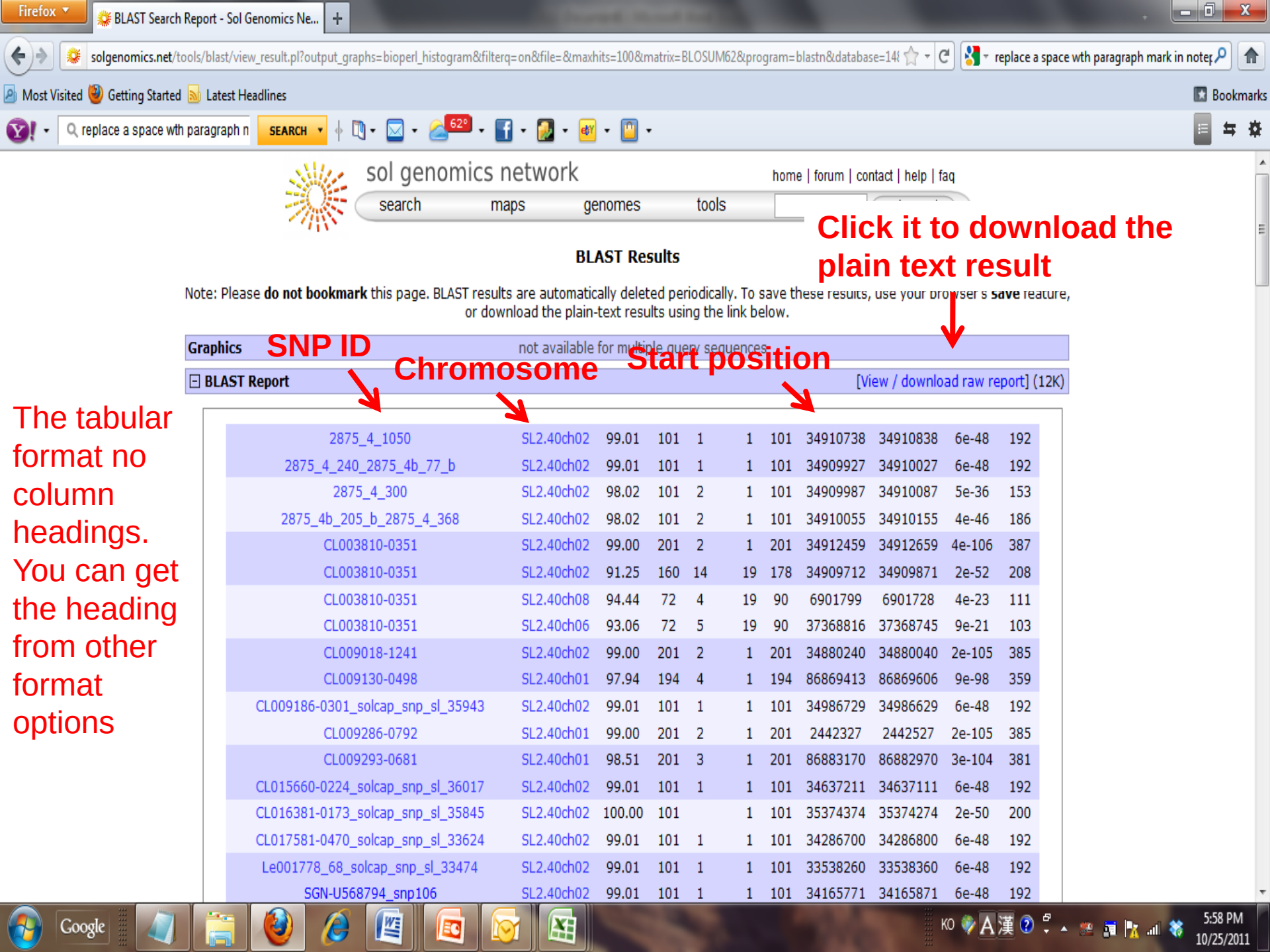

Firefox
Sol Genomics Network
http://solgenomics.net/
Google
Most Visited
Getting Started
Latest Headlines
Bookmarks
sgn
SEARCH
sol genomics network
home | forum | contact | help | faq
search
maps
genomes
tools
sol search
log in | new user
http://solgenomics.net/
BLAST
Sequence Analysis
BLAST
Alignment Analyzer
Tree Browser
Intron Finder
Mapping
Comparative Viewer
CAPS Designer
Seed BAC Finder
solQTL: QTL Mapping
Molecular Biology
Signal Peptide Finder
In Silico PCR
Systems Biology
SolCyc Biochemical Pathways
Bulk Query
Unigene and BAC information
FTP Site
ID Converter (SGN <=> TIGR)
Other
SGN Ontology Browser [beta]
Maps & Markers
CT233
CD15
C2_At4g15790
Breeders Toolbox
Genomes & Sequences
About SGN
News
Events
http://solgenomics.net/tools/blast/
EN
11:34 AM
9/26/2011


Firefox
SGN BLAST - Sol Genomics Network
http://solgenomics.net/tools/blast/index.pl
Most Visited
Getting Started
Latest Headlines
Bookmarks
sgn
SEARCH
sol genomics network
home | forum | contact | help | faq
search
maps
genomes
tools
sol search
log in | new user
NCBI BLAST
Simple
Advanced
Sequence Set
SGN Tomato Combined - WGS, BAC, and unigene sequences
db details
Program
BLASTN (nucleotide to nucleotide)
Advanced
Query sequence
single sequence only, use Advanced for multiple
Expect (e-value) Threshold
1e-10
Clear
Search
Substitution Matrix
BLOSUM62 (default)
Show Graphics
all
Max. hits to show
100
EN
11:38 AM
9/26/2011


Firefox
SGN BLAST - Sol Genomics Network
http://solgenomics.net/tools/blast/index.pl?db_id=&interface_type=1&preload_id=&flush_cache=&seq=&preload_type=
Most Visited
Getting Started
Latest Headlines
Bookmarks
sgn
SEARCH
sol genomics network
home | forum | contact | help | faq
search
maps
genomes
tools
sol search
log in | new user
NCBI BLAST
Simple
Advanced
This version of the BLAST online tool allows multiple query sequences, more control over running options, and more report formats.
Database (-d)
SGN Tomato Combined - WGS, BAC, and unigene sequences
db details
Program (-p)
BLASTN (nucleotide to nucleotide)
Query sequences (-i)
Go to: Tomato WGS Chromosome SL2.40
AND/OR upload multi-fasta query file
Browse...
Output format (-m)
0 - pairwise (default)
Substitution Matrix (-M)
BLOSUM62 (default)
Expectation value (-e)
1e-10
Max DB seqs to show hits
100
Google
EN
11:42 AM
9/26/2011


Firefox
SGN BLAST - Sol Genomics Network
http://solgenomics.net/tools/blast/index.pl?db_id=&interface_type=1&preload_id=&flush_cache=&seq=&preload_type=
Most Visited
Getting Started
Latest Headlines
Bookmarks
sgn
SEARCH
sol genomics network
home | forum | contact | help | faq
search
maps
genomes
tools
sol search
log in | new user
NCBI BLAST
Simple
Advanced
This version of the BLAST online tool allows multiple query sequences, more control over running options, and more report formats.
Database (-d)
SGN Tomato Combined - WGS, BAC, and unigene sequences
db details
Program (-p)
Query sequences (-i)
Tomato BAC end seqs - HindIII
Tomato BAC end seqs - MboI
Tomato BAC end seqs - Micro-Tom (HindIII)
Tomato BAC seqs (masked vs SGN UniRepeats)
Tomato BAC seqs (not repeat-masked)
Tomato Clone end seqs (BACs and Fosmids)
Tomato Fosmid end seqs
Tomato WGS Alternate Contigs (cabog1.00)
Tomato WGS Alternate Scaffolds (cabog1.00)
Tomato WGS Chromosomes (SL2.10)
Tomato WGS Chromosomes (SL2.30)
Tomato WGS Chromosomes (SL2.31)
Tomato WGS Chromosomes (SL2.40)
Tomato WGS Contigs (SL1.00)
Tomato WGS Contigs (SL1.03)
Tomato WGS Contigs (SL1.50)
Tomato WGS Contigs (SL2.10)
Tomato WGS Contigs (SL2.30)
Tomato WGS Contigs (SL2.31)
Tomato WGS Contigs (SL2.40)
Output format (-m)
Substitution Matrix (-M)
Expectation value (-e)
Max DB seqs to show hits
100
Tomato WGS Chromosome SL2.40
EN
11:43 AM
9/26/2011


Firefox
SGN BLAST - Sol Genomics Network
http://solgenomics.net/tools/blast/index.pl?db_id=&interface_type=1&preload_id=&flush_cache=&seq=&preload_type=
Most Visited
Getting Started
Latest Headlines
Bookmarks
sgn
SEARCH
This version of the BLAST online tool allows multiple query sequences, more control over running options, and more report formats.
Database (-d)
Tomato WGS Chromosomes (SL2.40)
db details
Program (-p)
BLASTN (nucleotide to nucleotide)
Query sequences (-i)
>1260_2_229
AAGTCCATTTCTGTCATTCTTGCAAGCTCATTGGCCATGGCACCACCTTCAGCTAGCAGAGGAAA
TTGAAAAAGCTTCCCTTTTTGACTTCAACTTGACACT
>1287_1_84c
TCCAACATCAGTCTTAATAATATTATCAGTTACAACCAGAGGCACAGCATTGAMGTTTCAGTTGC
TTTTGTTATGCATTTAACATTCCCAGTTTCATCCACA
>1287_1_93_1287_1_86_b
CAACATCAGTCTTAATAATATTATCAGTTACAACCAGAGGCACAGCATGAACGTTTCAGTTGCTT
TTGTTATGCATTTAACATTCCCAGTTTCATCCACAAA
AND/OR upload multi-fasta query file
Browse...
You can browse a query file to upload
Output format (-m)
8 - tabular
Substitution Matrix (-M)
Expectation value (-e)
Max DB seqs to show hits from (-b)
Filter query sequence (DUST with blastn, SEG with others) (-F)
0 - pairwise (default)
1 - query-anchored showing identities
2 - query-anchored no identities
3 - flat query-anchored, show identities
4 - flat query-anchored, no identities
5 - query-anchored no identities and blunt ends
6 - flat query-anchored, no identities and blunt ends
7 - XML Blast output
8 - tabular
9 - tabular with comment lines
10 - ASN, text
11 - ASN, binary
Tabular
Show Graphics
not available for multiple query seqs
all
Google
EN
10:38 PM
9/26/2011


Firefox
BLAST Search Report - Sol Genomics Ne...
solgenomics.net/tools/blast/view_result.pl?output_graphs=bioperl_histogram&filterq=on&file=&maxhits=100&matrix=BLOSUM62&program=blastn&database=14
replace a space wth paragraph mark in notep
Most Visited
Getting Started
Latest Headlines
Bookmarks
replace a space wth paragraph n
SEARCH
sol genomics network
home | forum | contact | help | faq
search
maps
genomes
tools
Click it to download the plain text result
BLAST Results
Note: Please do not bookmark this page. BLAST results are automatically deleted periodically. To save these results, use your browser's save feature, or download the plain-text results using the link below.
Graphics
not available for multiple query sequences
SNP ID
Chromosome
Start position
BLAST Report
[View / download raw report] (12K)
The tabular format no column headings. You can get the heading from other format options
2875_4_1050 SL2.40ch02 99.01 101 1 1 101 34910738 34910838 6e-48 192
2875_4_240_2875_4b_77_b SL2.40ch02 99.01 101 1 1 101 34909927 34910027 6e-48 192
2875_4_300 SL2.40ch02 98.02 101 2 1 101 34909987 34910087 5e-36 153
2875_4b_205_b_2875_4_368 SL2.40ch02 98.02 101 2 1 101 34910055 34910155 4e-46 186
CL003810-0351 SL2.40ch02 99.00 201 2 1 201 34912459 34912659 4e-106 387
CL003810-0351 SL2.40ch02 91.25 160 14 19 178 34909712 34909871 2e-52 208
CL003810-0351 SL2.40ch08 94.44 72 4 19 90 6901799 6901728 4e-23 111
CL003810-0351 SL2.40ch06 93.06 72 5 19 90 37368816 37368745 9e-21 103
CL009018-1241 SL2.40ch02 99.00 201 2 1 201 34880240 34880040 2e-105 385
CL009130-0498 SL2.40ch01 97.94 194 4 1 194 86869413 86869606 9e-98 359
CL009186-0301_solcap_snp_sl_35943 SL2.40ch02 99.01 101 1 1 101 34986729 34986629 6e-48 192
CL009286-0792 SL2.40ch01 99.00 201 2 1 201 2442327 2442527 2e-105 385
CL009293-0681 SL2.40ch01 98.51 201 3 1 201 86883170 86882970 3e-104 381
CL015660-0224_solcap_snp_sl_36017 SL2.40ch02 99.01 101 1 1 101 34637211 34637111 6e-48 192
CL016381-0173_solcap_snp_sl_35845 SL2.40ch02 100.00 101 1 101 35374374 35374274 2e-50 200
CL017581-0470_solcap_snp_sl_33624 SL2.40ch02 99.01 101 1 1 101 34286700 34286800 6e-48 192
Le001778_68_solcap_snp_sl_33474 SL2.40ch02 99.01 101 1 1 101 33538260 33538360 6e-48 192
SGN-U568794_snp106 SL2.40ch02 99.01 101 1 1 101 34165771 34165871 6e-48 192
Google
5:58 PM
10/25/2011

# Selection of Non-Redundant SNPs

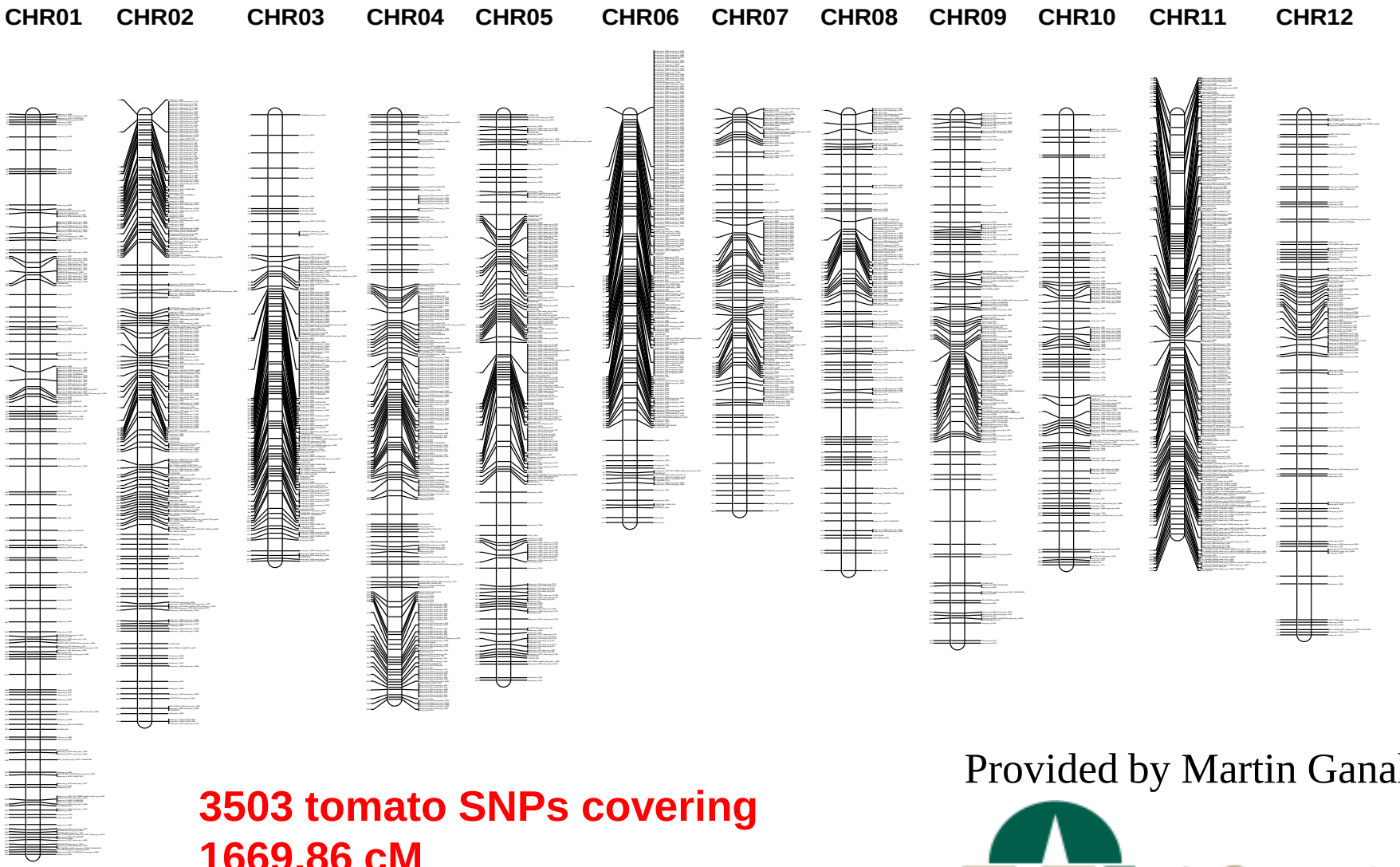
- Infinium array with a high density of SNPs is not necessary for all applications
- For mapping purposes, it may be more cost-effective to use a subset of SNPs covering the genome within a window that corresponds to expected recombination (e.g. 0.2 Mbp)
- Use physical or genetic map information for this selection
  - Two data sets: SNP data and physical map data
  - Combine these based on SNP ID using IF and COUNTIF functions

# Tomato Community Mapping Populations

- Processing (Proc) populations
  - Nested IBC (178 progeny; 1,130 SNPs)
  - Nested RIL (288 progeny; 1,232 SNPs)

- Fresh market (FM) populations
  - Fla.7776 x Fla.8383 $F_2$ (200 progeny; 657 SNPs)
  - NC33EB-1 x 091120-7 $F_2$ (195 progeny; 853 SNPs)
  - Fla. 7775 x NC1CELBR $F_2$ (250 progeny; 900 SNPs)

# Tomato Genetic Map of ExPen 2000




**3503 tomato SNPs covering**
**1669.86 cM**

Provided by Martin Ganal

TraitGenetics

# Two Subsets of SNPs

- We developed subsets of SNPs for Proc and FM germplasm using the genetic and physical map locations
- Selection criteria:
  - 0.2 cM or 0.2 Mbp windows
  - Evaluation results of flanking sequences for primer design via illumina ADT (Assay Design Tool)
  - Coverage of 12 chromosomes
  - Common SNPs between populations
  - Population specific SNPs

# 384 SNPs for the Processing Populations

| Group | # of SNP |
|---|---|
| Common IBC&RIL | 104 |
| IBC specific | 125 |
| RIL specific | 155 |
| IBC markers | 229 |
| RIL markers | 259 |

- IBC: Nested IBC (178 progeny)
- RIL: Nested RIL (288 progeny)

| Chromosome | # of SNP | |
|---|---|---|
| | IBC pop | RIL pop |
| Chr 1 | 14 | 13 |
| Chr 2 | 25 | 30 |
| Chr 3 | 33 | 16 |
| Chr 4 | 7 | 35 |
| Chr 5 | 44 | 76 |
| Chr 6 | 13 | 14 |
| Chr 7 | 9 | 11 |
| Chr 8 | 4 | 12 |
| Chr 9 | 14 | 16 |
| Chr 10 | 6 | 9 |
| Chr 11 | 51 | 16 |
| Chr 12 | 9 | 11 |
| Total | 229 | 259 |

# 384 SNPs for the Fresh Market Populations

| Group | # of SNP |
|---|---|
| Common among all three pops | 43 |
| Common btw SH&MM | 20 |
| Common btw SH&DP | 40 |
| Common btw MM&DP | 37 |
| SH specific | 79 |
| MM specific | 83 |
| DP specific | 82 |
| SH markers | 182 |
| MM markers | 183 |
| DP markers | 202 |

- SH: Fla.7776 x Fla.8383 (200 $F_2$)
- MM: NC33EB-1 x 091120-7 (195 $F_2$)
- DP: Fla.7775 x NC1CELBR (250 $F_2$)

| Chromosome | # of SNP | | |
|---|---|---|---|
| | SH pop | MM pop | DP pop |
| Chr 1 | 14 | 16 | 8 |
| Chr 2 | 8 | 8 | 9 |
| Chr 3 | 9 | 16 | 15 |
| Chr 4 | 38 | 43 | 39 |
| Chr 5 | 20 | 26 | 13 |
| Chr 6 | 9 | 7 | 9 |
| Chr 7 | 9 | 5 | 9 |
| Chr 8 | 10 | 5 | 15 |
| Chr 9 | 7 | 19 | 24 |
| Chr 10 | 14 | 5 | 17 |
| Chr 11 | 9 | 9 | 8 |
| Chr 12 | 35 | 24 | 36 |
| Total | 182 | 183 | 202 |

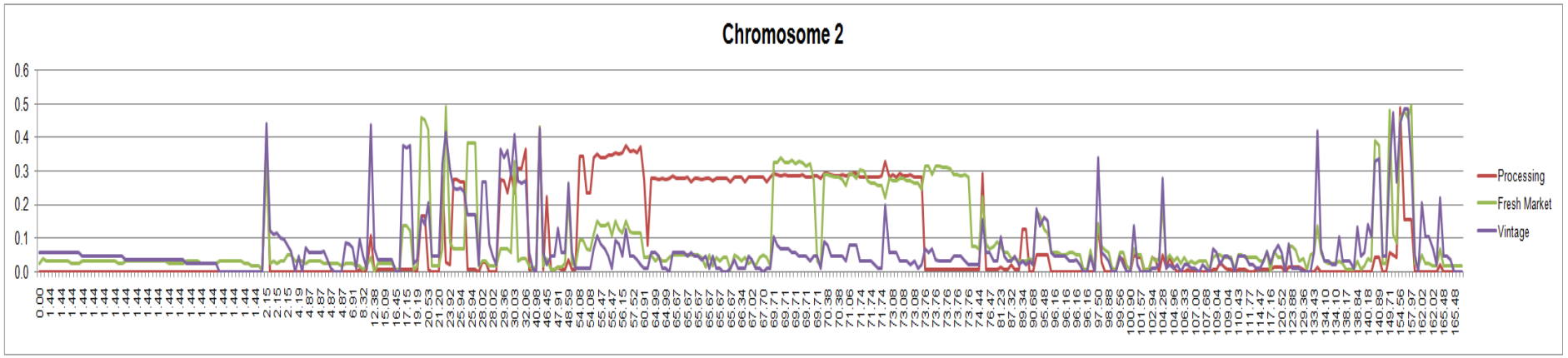
# Genotyping with the Subsets of SNPs

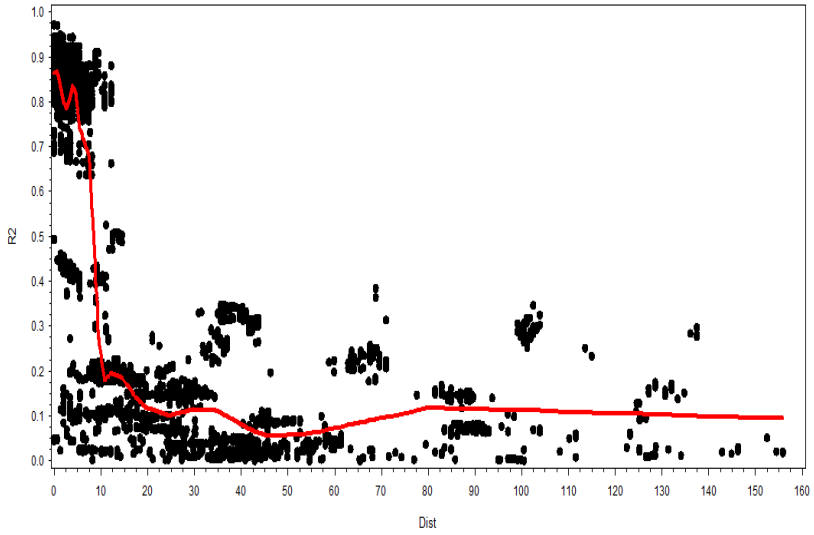
- Two 384 SNP lists with flanking sequences are available on the workshop URL (http://www.extension.org/pages/61007)
  - FM_384SNPs_SolCAP_10032011.csv
  - Proc_384SNPs_SolCAP_10122011.csv
- Genotyping platforms:
  - BeadXpress assay at OSU: Proc populations
  - KBioscience (www.kbioscience.com): FM populations

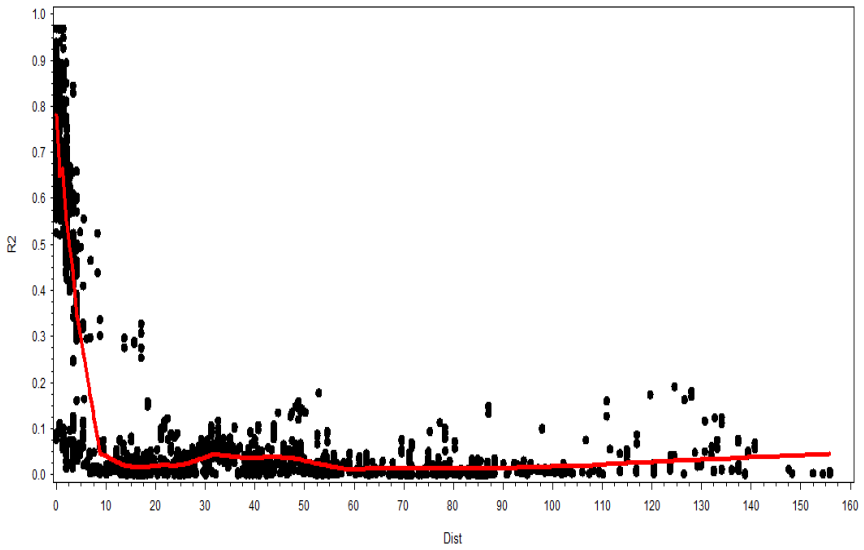
# Minor Allele Frequency (MAF)

- Frequency of the less common allele in a given population
- Visualize allelic variation between market classes
- SNPs with a minor allele frequency of 10% or greater are commonly used for LD analysis (SNPs with low levels of MAF can result in inaccurate estimates)
- MAF >10% (depending on pop. Size) for Assoc. analysis
- Use the SNP dataset with ‘AB’ code to calculate MAF in spreadsheet

# Distribution of MAF along chromosome 2 for different market classes




## Patterns of LD decay for Chr 2 in Proc and FM germplasm

# Summary

- SolCAP genotyped 489 tomato accessions using 7,720 SNPs (7,627 SNPs with good calls).
- The cluster file (v1) for the tomato Infinium array is released to public with the Data Access Agreement.
- The use of a subset of SNPs may be more cost-effective for some applications (e.g. genetic mapping)
- Two subsets of SNPs (n=384) were developed for the Proc and FM community mapping populations.
- MAF and LD analysis will lead to develop an efficient strategy for tomato improvement using genomic tools

# Acknowledgments


**OSU**

David Francis
Heather Merk
Troy Aldrich
Nancy Huarachi
Daniel Thomas
Caleb Orchard

**UCD**

Allen Van Deynze
Kevin Stoffel
Alex Kozic

**MSU**

David Douches
Robin Buell
John Hamilton
Dan Zarka
Kelly Zarka

**Cornell**

Walter De Jong
Lucas Mueller

**INRI**

Mathilde Causse

**Trait Genetics**

Martin Ganal
Gregor Durstewitz

**Illumina**

Cindy Lawley
Maila Crist

**Funding:** USDA NIFA United States Department of Agriculture National Institute of Food and Agriculture