Variant analysis of Arabidopsis EMS pools
A full GATK4 command Line workflow

Stéphane Plaisance [VIB - Nucleomics Core, nucleomics@vib.be]

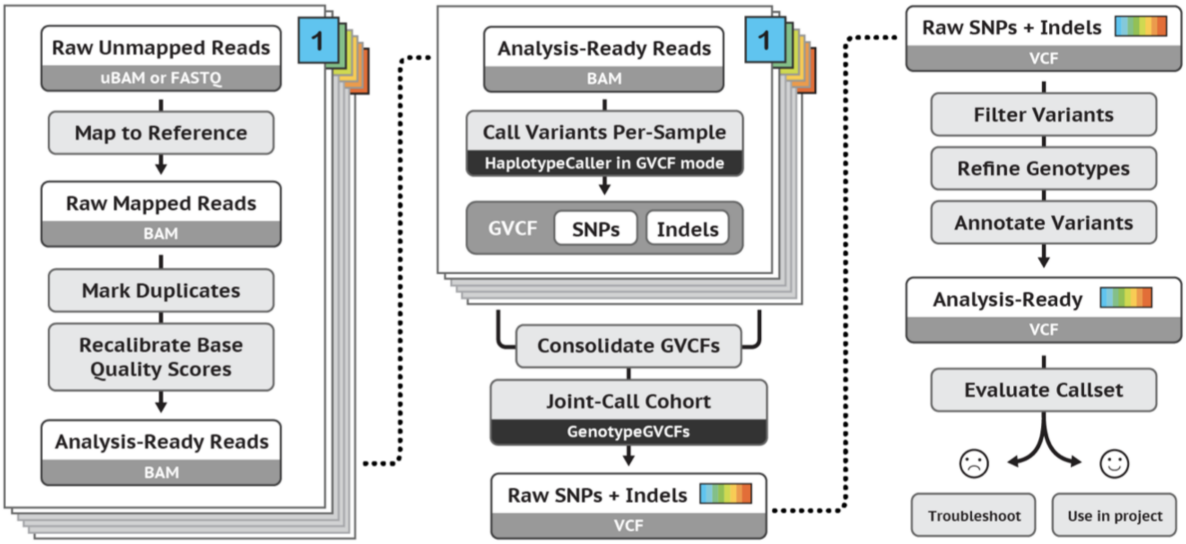
April 20th, 2020 - version 1.0

Contents

last edits: Wed May 06, 2020 v1.0

Forewords

This report details the variant analysis of public Arabidopsis WGS data. The aim of this report is to guide users through the process of processing read data through the current GATK4 **Germline-short-variant-discovery-SNPs-Indels workflow** until obtaining a list of calibrated variants ready for candidate selection. Variant annotation is then done using **SnpEff** and examples of **SnpSift** filtering are provided to guide further data exploration by the user.



It is likely that the tools used here change in the near future (new versions of GATK) and we therefore cannot guaranty that the presented code will run without errors in the future. Please refer to the GATK support forum if this would be the case.

The data was obtained from the NCBI SRA repository and is not yet associated to a publication which would have been nice to see some validation of the results. We chose this dataset because it resembles a project submitted to us by a customer and aimed at finding variants causative for a phenotype observed after EMS (ethyl-methane sulfonate) mutagenesis of Drosophila flies and their back-crossing to the parental line. A nice review of such project can be read in a 2015 review (Thole and Strader 2015).

Material & Method

Software requirements

In order to reproduce this workflow, the user will need a Unix server with

- sufficient amounts of RAM (at least 4GB, more is better, adjust in the code)
- multiple CPU for mapping the reads faster (at least 4, more is better, adjust in the code)
- wget, gawk (use the package manager to install)
- samtools (get from: http://www.htslib.org/download/) (Li et al. 2009)
- bwa (get from: https://github.com/lh3/bwa) (Li and Durbin 2010)
- java JRE 1.8 or higher to run GATK and SnpEff
- GATK4 (get the latest version from: https://github.com/broadinstitute/gatk/releases) (Poplin et al. 2017)
- SnpEff (get from: http://snpeff.sourceforge.net/download.html) (Cingolani et al. 2012)

Each of the packages above has a online detailed documentation that should be read in order to learn using these tools correctly. Users are strongly advised to register to the GATK4 user forum (https://gatk.broadinstitute.org/hc/en-us/community/topics) where they can submit questions raised during the analysis and ask for advise.

SRA Data

The selected SRA dataset PRJNA574113 groups 5 plant pools obtained from the whole genome sequences of two suppressor families, their non-suppressed siblings, and the hpat1/3 background genotype (**Col-0**). The samples were sequenced on a *Illumina HiSeq 4000* instrument (all PE-150 reads except for one sample in SE-150).

Whole genome sequencing of two independent hpat1/3 suppressor families Accession: PRJNA574113 ID: 574113

To learn more about the molecular components involved in regulating pollen tube growth in Arabidopsis, we performed a suppressor screen and identified several plants that exhibited robust suppression of hpat1/3 pollen fertility defects. After several generations of backcrossing, we sequenced the genomes of two suppressor families and identified candidate suppression-causing mutations for each suppressor family. This project contains the whole genome sequences of two suppressor families, their non-suppressed siblings, and the hpat1/3 background genotype.
Less...

Accession	PRJNA574113
Data Type	Raw sequence reads
Scope	Multispecies
Submission	Registration date: 25-Sep-2019 **University of Michigan**
Relevance	Evolution

The read files associated with PRJNA574113 were downloaded manually from SRA after failing to obtain correctly formatted reads using the NBCI toolbox (read names were incorrect and did not contain the flow-cell coordinate required to find duplicates). The following code section will allow users download the reads and process them in a similar way to ours.

The sequencing information for one of the pools (*FRH2 nonsup*) is shown next.

SRX6899729: Whole genome seq of Arabidopsis thaliana: leaf
1 ILLUMINA (Illumina HiSeq 4000) run: 55.9M spots, 16.9G bases, 4.8Gb downloads

Design: DNA was extracted from small, young leaves using the DNeasy Mini kit from Qiagen. DNA was pooled together from ~10-20 plants per genotype/sample.

Submitted by: University of Michigan

Study: Whole genome sequencing of two independent hpat1/3 suppressor families
PRJNA574113 • SRP223178 • All experiments • All runs
hide Abstract
To learn more about the molecular components involved in regulating pollen tube growth in Arabidopsis, we performed a suppressor screen and identified several plants that exhibited robust suppression of hpat1/3 pollen fertility defects. After several generations of backcrossing, we sequenced the genomes of two suppressor families and identified candidate suppression-causing mutations for each suppressor family. This project contains the whole genome sequences of two suppressor families, their non-suppressed siblings, and the hpat1/3 background genotype.

Sample:
SAMN12840255 • SRS5431925 • All experiments • All runs
Organism: Arabidopsis thaliana

Library:
Name: 127850
Instrument: Illumina HiSeq 4000
Strategy: WGS
Source: GENOMIC
Selection: unspecified
Layout: PAIRED

Runs: 1 run, 55.9M spots, 16.9G bases, 4.8Gb

Run	# of Spots	# of Bases	Size	Published
SRR10178322	55,907,035	16.9G	4.8Gb	2019-09-26

ID: 9077088

Reference data

The Arabidopsis reference genome (Col-0), gene annotations, and known variants were obtained from different repositories and prepared to be used in the GATK analysis and later variant annotation. The code below should allow producing the same reference used in this work (as long as the web links remain valid).

GATK4 Analysis

Read mapping with BWA mem

A configuration file is produced in order to loop through the read files and produce mappings with correct names and metadata information.

The reads files are aligned to the reference genome and alignments are produced.

The samtools flagstats table for the five sample alignment results is shown next.

results	SRR10178322	SRR10178323	SRR10178324	SRR10178325	SRR10178326
total (QC-passed reads + QC-failed reads)	112080396	82449522	31960505	39730973	33097524
secondary	266326	148718	42122	123297	106350
supplementary	0	0	0	0	0
duplicates	0	0	0	0	0
mapped	111790803	82193374	31383145	39325198	32834020
mapped %	99.74%	99.69%	98.19%	98.98%	99.20%
paired in sequencing	111814070	82300804	0	39607676	32991174
read1	55907035	41150402	0	19803838	16495587
read2	55907035	41150402	0	19803838	16495587
properly paired	109487370	80580446	0	38268188	31718770
properly paired %	97.92%	97.91%	N/A	96.62%	96.14%
with itself and mate mapped	111441256	81960218	0	39140280	32655890
singletons	83221	84438	0	61621	71780
singletons %	0.07%	0.10%	N/A	0.16%	0.22%
with mate mapped to a different chr	1261494	927784	0	556242	644610
with mate mapped to a different chr (mapQ>=5)	749828	539817	0	324283	427179

GATK Cleanup & MarkDuplicates

Read are sorted by queryname in order to find duplicates with MarkDuplicates. Duplicates are not physically removed but instead 'flagged' and this extra info will be used to count them accordingly during calling.

The MarkDuplicate counts for the five samples are reported in the next table.

IDS	SRR10178322	SRR10178323	SRR10178324	SRR10178325	SRR10178326
LIBRARY	SRS5431925	SRS5431924	SRS5431923	SRS5431922	SRS5431921
UNPAIRED_READS_EXAMINED	83221	84438	31341023	61621	71780
READ_PAIRS_EXAMINED	55720628	40980109	0	19570140	16327945
SECONDARY_OR_SUPPLEMENTARY_RDS	266326	148718	42122	123297	106350
UNMAPPED_READS	289593	256148	577360	405775	263504
UNPAIRED_READ_DUPLICATES	54578	53068	8851169	24964	29739
READ_PAIR_DUPLICATES	9556877	6216610	0	2548373	1633729
READ_PAIR_OPTICAL_DUPLICATES	2952702	2293091	0	1417805	870658
PERCENT_DUPLICATION	0.171876	0.152189	0.282415	0.13065	0.100746
ESTIMATED_LIBRARY_SIZE	192827182	177606225		139610660	151360701

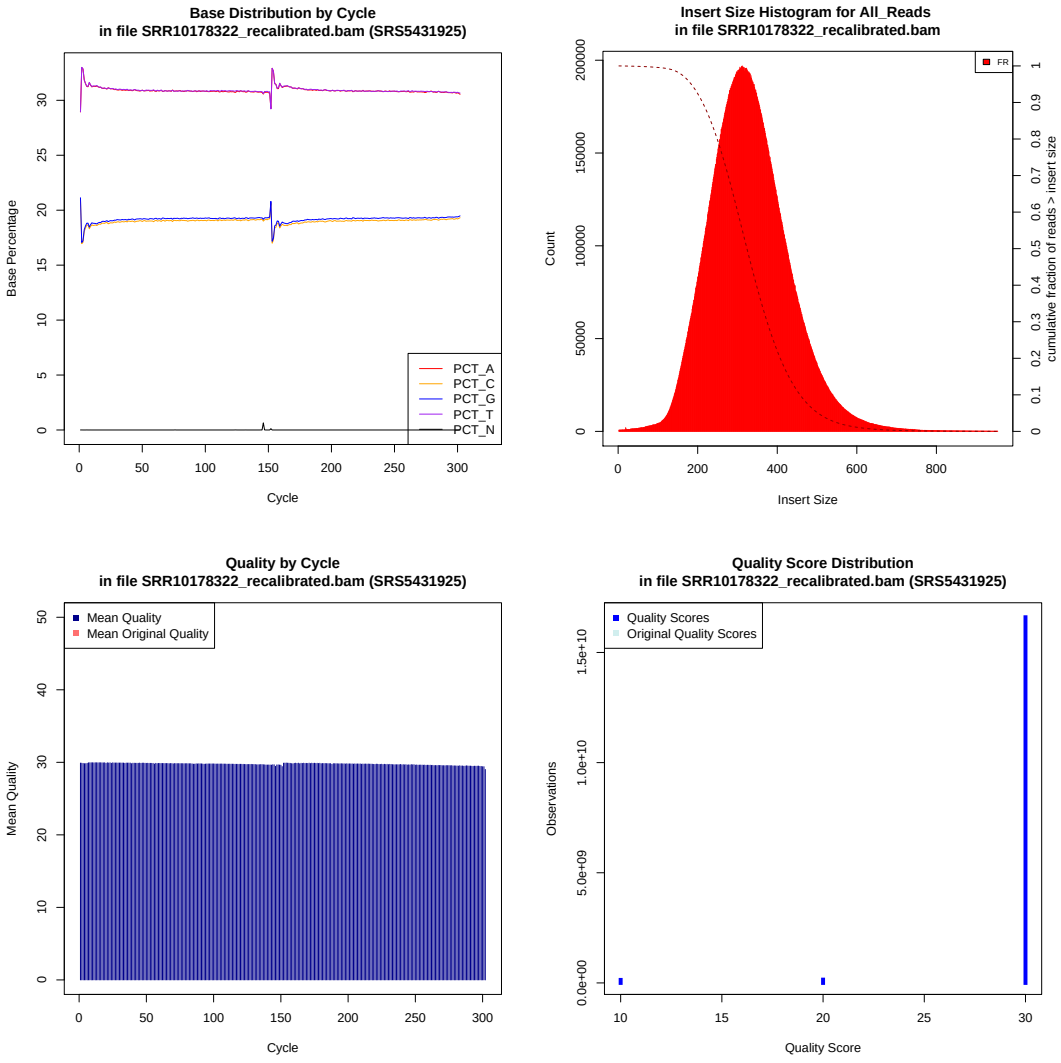
The *_mrkdup_srt-tags.bam_ValidateSamFile.txt* files all report 'No errors found', confirming that the obtained BAM format is compliant.

NB: The quantity of duplicates is relatively large in this experiment with ranges between 10% and 28% (SE-150 parental sample) when taking into account both duplicate types. This seems to suggest that the library was over-amplified and/or loaded at high titre. The GATK pipeline will correct for this and reads having been marked will be counted only once during variant calling.

GATK basecall quality score recalibration (BQSR)

Basecalling is often biased, known variants found in a samplle can be con,sidered as true and used to improve the basecalling quality score distribution and thereby create a more suitable mapping data for variant calling. This is done next using the **BaseRecalibrator** function of the GATK pipeline.

The recalibration identifies potential biases and measures the insert size based on the read pair mappings. Results are shown for one sample (SRR10178322: tFRH2_nonsup).



GATK4 HaplotypeCaller for Germline short variant discovery (SNPs + Indels)

This workflow uses the GATK4 **HaplotypeCaller** module to identify short variants based on the alignments. This is the state of the art method in today's pipeline. Each genome is called separately and the results are sacved in **gVCF** format which is looseless as compared to calls in VCF format where only genome positions with differences as compared to the reference are recorded. The gVCF data is larger than VCF and needs merging and conversion before we can review and use it (done next).

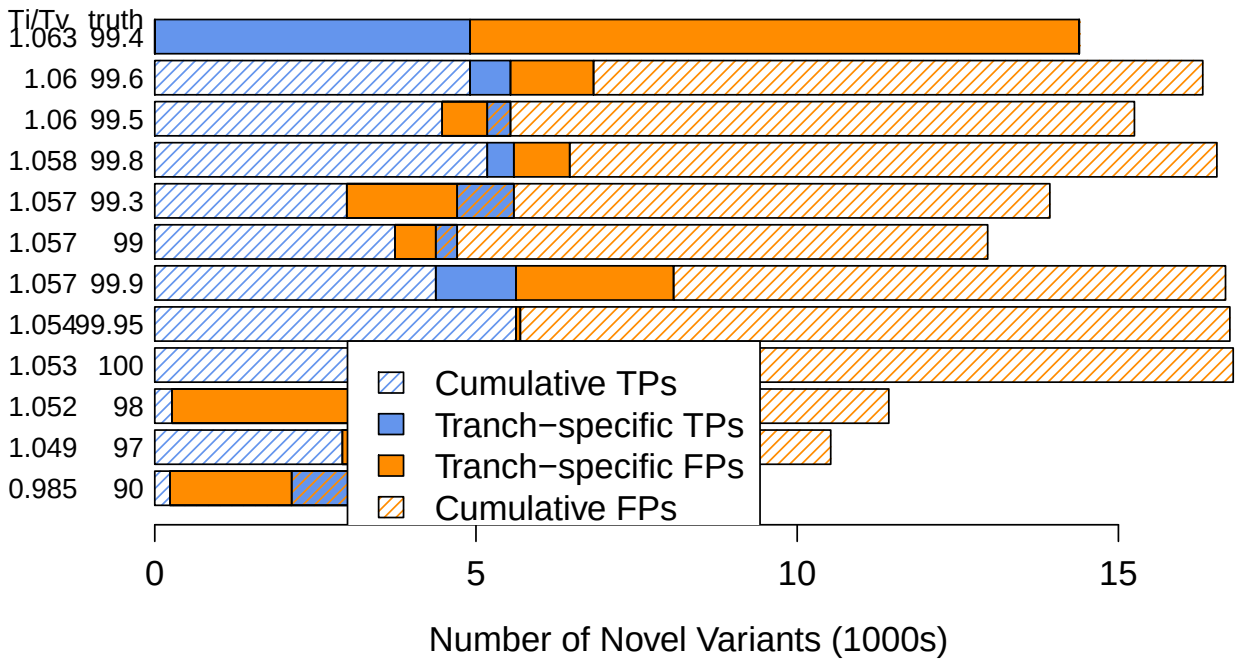
GATK Merge multiple gVCF and convert to VCF

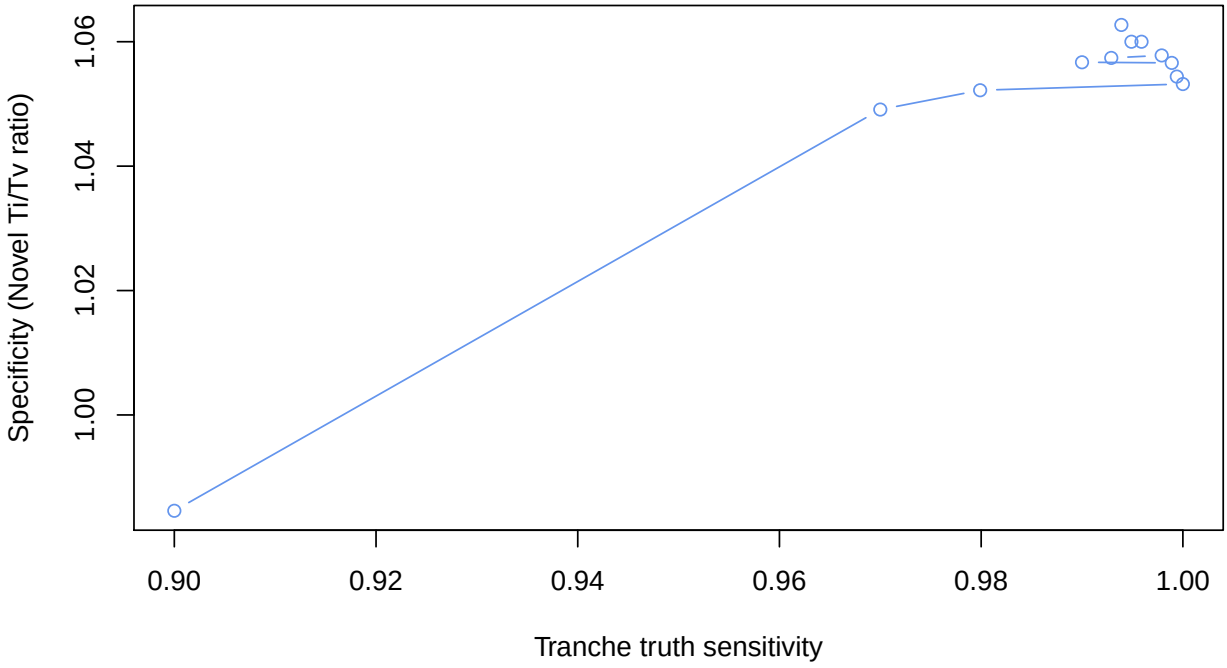
The five gVCF resulting files are merged and the merge converted to a classical VCF file. Thanks to the full information containd in th gVCF, genomes where a variant was not measured (no-calls) are also reported as such instead of as reference-calls. Such no-calls could be false negative variant calls and are important to identify for later screening and filtering.

GATK4 variant quality score recalibration (VQSR)

Like for basecalling scores previously, variant calling is biased buy multiple factors. Using the similar logic applied to basecalls, known variants found in one or more genomes are considered as true positives and their measured variant calling score can be analysed to produce a statistical model used to recalibrate all variant scores in the dataset. This is done during the VQSR part of the pipeline and using the known variants from the EnsEMBL database of the 1001 Ath genomes.

The plots derived from the pipeline are shown next to illustrrate the complex recalibration process.




Specificity (Novel Ti/Tv ratio)
1.00
1.02
1.04
1.06
0.90
0.92
0.94
0.96
0.98
1.00
Tranche truth sensitivity

SnpEFF variant annotation and filtering

The final result of the GATK pipeline is a multi-genome VCF file with shared annotations and all 5 genotypes represented in a separate column. The variants are still anonymous and their effect is still unknown. To add this extra information, we chose to use the popular **SnpEff** software and its **SnpSift** filtering counterpart.

The code below illustrates the annotation and filtering processes with example commands for candidate filtering to be extended by the end user.

A very comprehensive documentation for SnpEff and SnpSift can be found on http://snpeff.sourceforge.net/index.html. We strongly suggest to fully read the SnpEFF documentation in order to discover the many tools and methods that can be found in the package.

SnpEff on the VQSR VCF data

The main results of the SNPEff annotation process are shown next. The report contains much more detailed information and only key figures are reproduced here.

Variants rate details

Chromosome	Length	Variants	Variants rate
1	30,427,671	10,658	2,854
2	19,698,289	6,922	2,845
3	23,459,830	11,133	2,107
4	18,585,056	8,890	2,090
5	26,975,502	9,490	2,842
Mt	366,924	319	1,150
Pt	154,478	21	7,356
Total	**119,667,750**	**47,433**	**2,522**

Number variants by type

Type	Total
SNP	25,260
MNP	0
INS	8,859
DEL	13,314
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	**47,433**

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	512	0.285%
LOW	1,035	0.577%
MODERATE	1,057	0.589%
MODIFIER	176,751	98.548%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	947	64.819%
NONSENSE	26	1.78%
SILENT	488	33.402%

Missense / Silent ratio: 1.9406

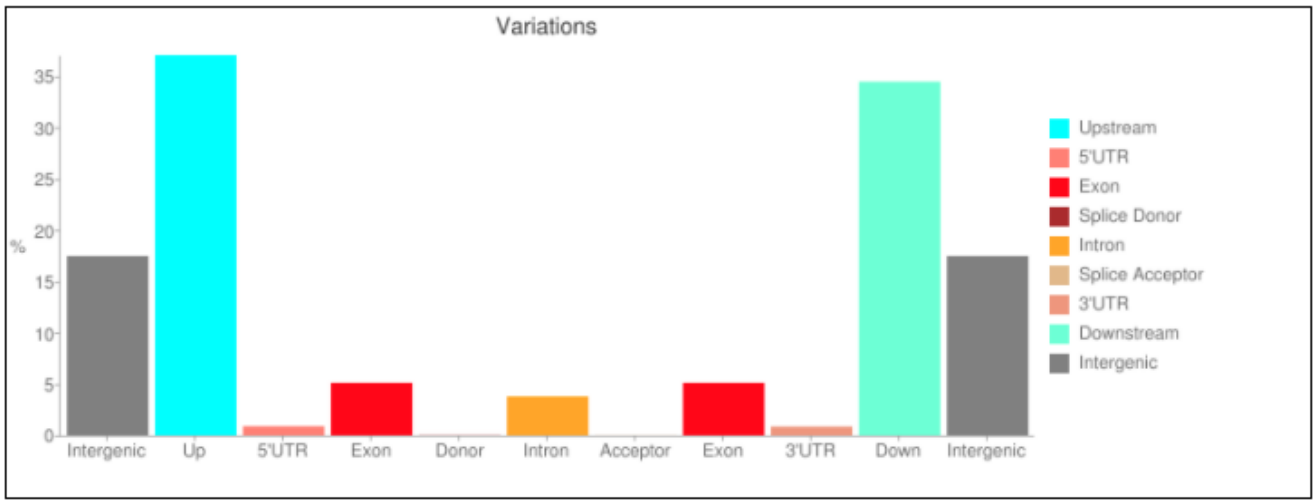
Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	1,521	0.845%
5_prime_UTR_premature_start_codon_gain_variant	22	0.012%
5_prime_UTR_variant	1,570	0.872%
conservative_inframe_deletion	26	0.014%
conservative_inframe_insertion	18	0.01%
disruptive_inframe_deletion	52	0.029%
disruptive_inframe_insertion	28	0.016%
downstream_gene_variant	61,844	34.352%
exon_loss_variant	4	0.002%
frameshift_variant	392	0.218%
initiator_codon_variant	1	0.001%
intergenic_region	31,354	17.416%
intragenic_variant	2	0.001%
intron_variant	7,285	4.047%
missense_variant	941	0.523%
non_coding_transcript_exon_variant	7,227	4.014%
non_coding_transcript_variant	5	0.003%
splice_acceptor_variant	31	0.017%
splice_donor_variant	56	0.031%
splice_region_variant	599	0.333%
start_lost	49	0.027%
stop_gained	36	0.02%
stop_lost	9	0.005%
stop_retained_variant	1	0.001%
synonymous_variant	487	0.271%
upstream_gene_variant	66,468	36.921%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	61,844	34.481%
EXON	9,170	5.113%
INTERGENIC	31,354	17.482%
INTRON	6,786	3.784%
SPLICE_SITE_ACCEPTOR	27	0.015%
SPLICE_SITE_DONOR	53	0.03%
SPLICE_SITE_REGION	533	0.297%
TRANSCRIPT	7	0.004%
UPSTREAM	66,468	37.059%
UTR_3_PRIME	1,521	0.848%
UTR_5_PRIME	1,592	0.888%



SnpEff add extra annotations

More annotations can be added based on teh variant distribution and phenotype/genotype expectations. The examples provided are not exhaustive but more incentive to help the user define efficient filters and apply them to the data.

SnpSift primer to filter variant candidates

Now that genome annotations were added, they can be used to reduce the very long list of variants to a shorter list(s) of candidate variants, taking into account their expected metrics and distribution. One filtering step will not lead to the final list and the user will need to be inventive and spamart to define those parameters that define the ideal variant and try many combinations.

Some filtering examples are reproduced here after converting the filtered VCF data to more readable tables with **SnpSift extractFields**.

REM: tables below have been cut to fit in the page width.

extractFields_01.txt (first 5 rows)

CHROM	POS	ID	AF	REF	ALT
1	30872	.	0.2	C	T
1	88937	.	1.0	G	A
1	99686	.	0.2	C	T
1	926694	ENSVATH01009781	1.0	C	T
1	1086696	.	0.1	C	T

ANN[*].GENE	ANN[*].HGVS_P
MIR838A,PPA1,LHY,LHY,LHY,LHY,LHY,DCL1,DCL1	.,.,.,.,.,.,.,.,.
AT1G01210,CYP78A8,RABA3,FKGP	.,.,.,.
AT1G01240,AT1G01240,AT1G01240,FKGP,AT1G01225,AT1G01230,ERF023,AT1G01230-AT1G01240	.,.,.,.,.,.,.,.
AT1G03720,AT1G03700,AT1G03710,AT1G03730,AT1G03710,AT1G03720-AT1G03730	.,.,.,.,.,.
AT1G04150,XI-B	.,.

extractFields_02.txt (first 5 rows)

CHROM	POS	ID	AF	REF	ALT
1	30872	.	0.2	C	T
1	88937	.	1.0	G	A
1	99686	.	0.2	C	T
1	926694	ENSVATH01009781	1.0	C	T
1	1086696	.	0.1	C	T

ANN[*].GENE
MIR838A,PPA1,LHY,LHY,LHY,LHY,LHY,DCL1,DCL1
AT1G01210,CYP78A8,RABA3,FKGP
AT1G01240,AT1G01240,AT1G01240,FKGP,AT1G01225,AT1G01230,ERF023,AT1G01230-AT1G01240
AT1G03720,AT1G03700,AT1G03710,AT1G03730,AT1G03710,AT1G03720-AT1G03730
AT1G04150,XI-B

ANN[*].HGVS_P	CasesFRH1_sup	CasesFRH2_sup
.,.,.,.,.,.,.,.,.	0,0,0	0,1,1
.,.,.,.	1,0,2	1,0,2
.,.,.,.,.,.,.,.	0,1,1	0,0,0
.,.,.,.,.,.	1,0,2	0,0,0
.,.	0,1,1	0,0,0

extractFields_03.txt (first 5 rows)

CHROM	POS	ID	AF	REF	ALT	GEN[*].GT	ANN[0].GENEID
1	30872	.	0.2	C	T	0/0,0/0,0/1,0/1,0/0	AT1G01046
1	88937	.	1.0	G	A	1/1,1/1,1/1,1/1,1/1	AT1G01210
1	99686	.	0.2	C	T	0	0,0
1	926694	ENSVATH01009781	1.0	C	T	./.,1/1,1/1,./.,1/1	AT1G03720
1	1086696	.	0.1	C	T	0/0,0	1,0/0,0/0,0/0

ANN[*].IMPACT
MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
MODIFIER,MODIFIER,MODIFIER,MODIFIER
MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
MODIFIER,MODIFIER

ANN[*].HGVS_P	CasesFRH1_sup	CasesFRH2_sup	LOF
.,.,.,.,.,.,.,.,.	0,0,0	0,1,1	.
.,.,.,.,.	1,0,2	1,0,2	.
.,.,.,.,.,.,.,.	0,1,1	0,0,0	.
.,.,.,.,.,.,.	1,0,2	0,0,0	.
.,.	0,1,1	0,0,0	.

extractFields_04.txt (first 5 rows)

CHROM	POS	ID	AF	REF	ALT	GEN[*].GT
1	30872	.	0.2	C	T	0/0,0/0,0/1,0/1,0/0
1	88937	.	1.0	G	A	1/1,1/1,1/1,1/1,1/1
1	99686	.	0.2	C	T	0
1	926694	ENSVATH01009781	1.0	C	T	./.,1/1,1/1,./.,1/1
1	1086696	.	0.1	C	T	0/0,0

ANN[0].GENEID	ANN[*].IMPACT
AT1G01046	MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
AT1G01210	MODIFIER,MODIFIER,MODIFIER,MODIFIER
AT1G01240	MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
AT1G03720	MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER,MODIFIER
AT1G04150	MODIFIER,MODIFIER

ANN[*].HGVS_P	CasesFRH1_sup	CasesFRH2_sup	ControlsFRH1_sup	ControlsFRH2_sup	LOF
.,.,.,.,.,.,.,.,.	0,0,0	0,1,1	0,2,2	0,1,1	.
.,.,.,.,.	1,0,2	1,0,2	4,0,8	4,0,8	.
.,.,.,.,.,.,.,.	0,1,1	0,0,0	0,1,1	0,2,2	.
.,.,.,.,.,.,.	1,0,2	0,0,0	2,0,4	3,0,6	.
.,.	0,1,1	0,0,0	0,0,0	0,1,1	.

Further filter and identify EMS candidates

EMS is expected to produce **G:C=>A:T** SNPs in a majority of cases. This is used next to filter suppressor-specific loci displaying the expected base substitution.

Filtering for HIGH impact with SnpSift

The High impact variant introduce premature stop codons or frame-shifts in the transcript and constitute the low hanging fruit for variant analysis.

The table below returns the effect of filtering by genotype + EMS filter, then adding **HIGH impact** requirement.

	count	extracted in
EMS candidates	7695	extractFields_04.txt
FRH1_sup variants	621	extractFields_05.txt
FRH2_sup variants	325	extractFields_06.txt
FRH1_sup variants with HIGH impact	1	extractFields_07.txt
FRH2_sup variants with HIGH impact	4	extractFields_08.txt

Candidate EMS variants with high impact for FRH1_sup

extractFields_07.txt: FRH1_sup candidate with high impact

CHROM	POS	ID	AF	REF	ALT	GEN[*].GT	ANN[0].GENEID
3	14586470	ENSVATH08093138	0.1	G	A	0/0,0	1,0

ANN[*].IMPACT	ANN[*].HGVS_P	CasesFRH1_sup	CasesFRH2_sup	ControlsFRH1_sup
HIGH,MODIFIER,MODIFIER	.,.,.	0,1,1	0,0,0	0,0,0

ControlsFRH2_sup	LOF
0,1,1	.

The candidate is a known ensembl 1001 genomes variant ENSVATH08093138 in the **AT3G42450** gene that codes for a **transposable_element_gene;similar to zinc knuckle (CCHC-type) family protein** with no apparent protein expression link.

The plant Ensembl variant page returns:

Candidate EMS variants with high impact for FRH2_sup

extractFields_08.txt: FRH2_sup candidates with high impact

CHROM	POS	ID	AF	REF	ALT	GEN[*].GT
3	15142729	ENSVATH02373908	0.1	C	T	0/0,0/0,0/0,0
3	15142732	ENSVATH14289901	0.1	C	T	0/0,0/0,0/0,0
3	20935829	.	0.2	C	T	0/0,0/0,0/0,1/1,0/0
3	22051051	.	0.1	C	T	0/0,0/0,0/0,0/1,0/0

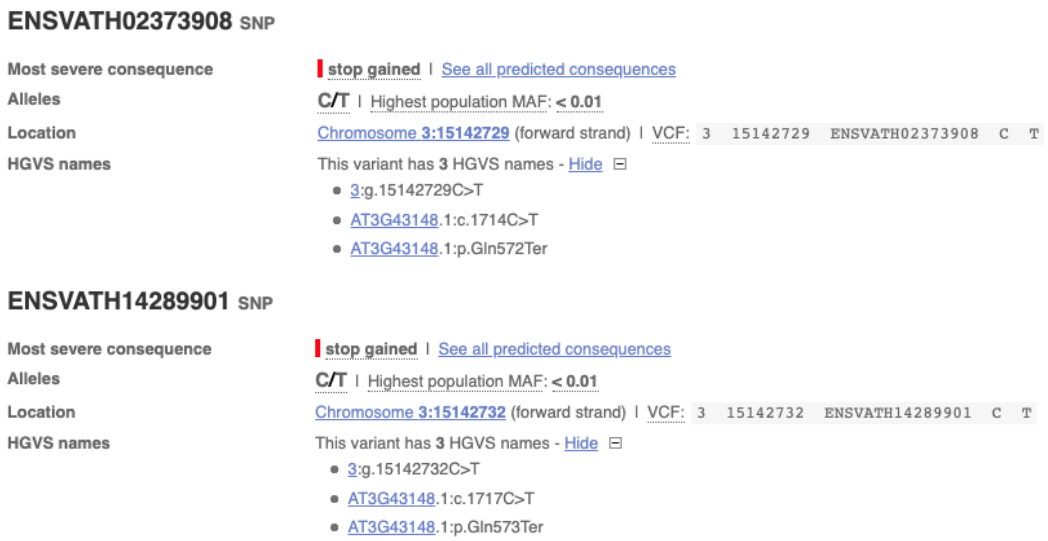
ANN[0].GENEID	ANN[*].IMPACT	ANN[*].HGVS_P
AT3G43148	HIGH,MODIFIER,MODIFIER,MODIFIER,MODIFIER	p.Gln572*,.,.,.,.
AT3G43148	HIGH,MODIFIER,MODIFIER,MODIFIER,MODIFIER	p.Gln573*,.,.,.,.
AT3G56470	HIGH,MODIFIER,MODIFIER,MODIFIER	p.Gln306*,.,.,.
AT3G59690	HIGH,MODIFIER,MODIFIER,MODIFIER	p.Gln376*,.,.,.

CasesFRH1_sup	CasesFRH2_sup	ControlsFRH1_sup	ControlsFRH2_sup
0,0,0	0,1,1	0,1,1	0,0,0
0,0,0	0,1,1	0,1,1	0,0,0
0,0,0	1,0,2	1,0,2	0,0,0
0,0,0	0,1,1	0,1,1	0,0,0

LOF
(AT3G43148|AT3G43148|1|1.00)
(AT3G43148|AT3G43148|1|1.00)
.
.

- two LOF **AT3G43148** variants predicted for FRH2 are known 1001 genome variants in a 673aa **myosin heavy chain-like protein** F4IXU3-1
  - ENSVATH02373908
  - ENSVATH14289901
- the **AT3G56470** gene codes for a 367 aa **F-box containing** protein Q9LXZ3
- the **AT3G59690** gene codes for a 517 aa **IQ-domain 13** plasma membrane protein Q9M199

The plant Ensembl variant pages returns:

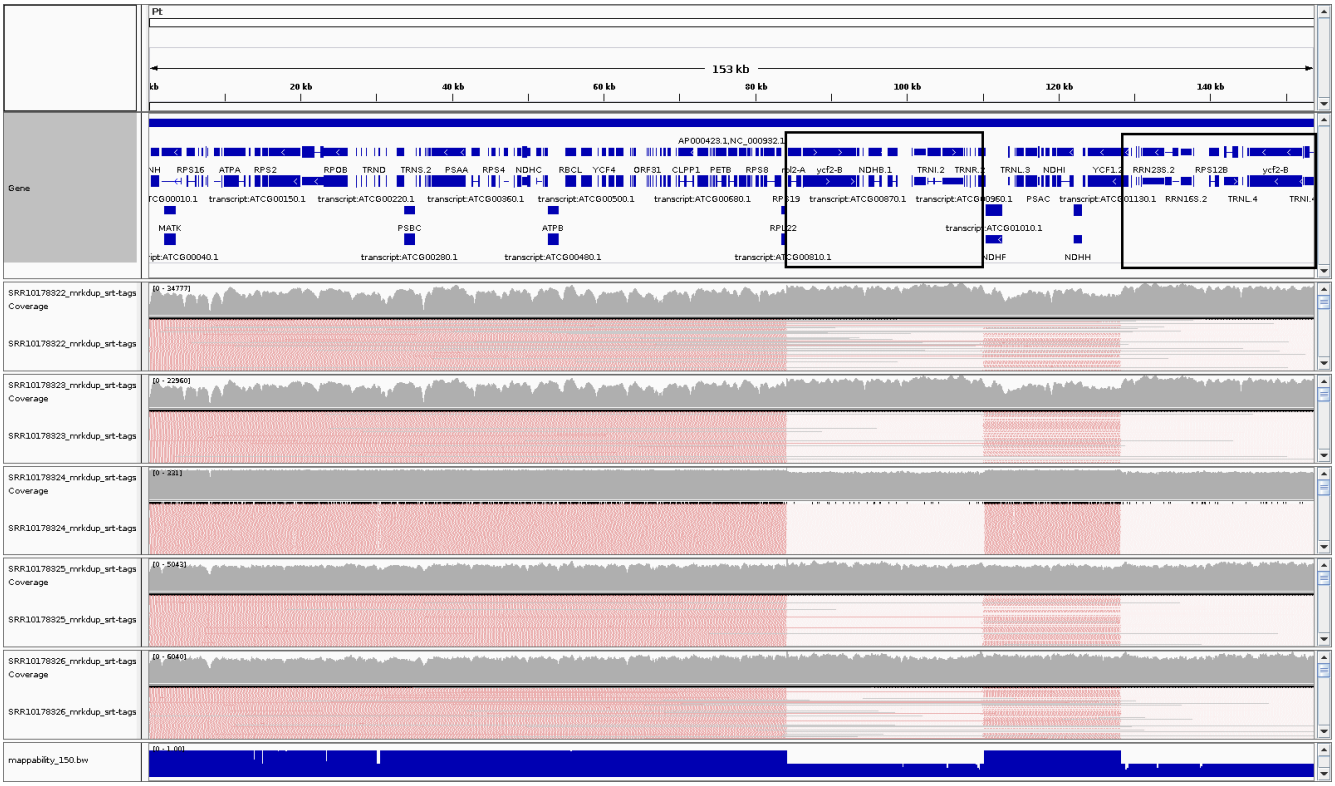


Many other filters can be built in order to define relevant subset for your own needs and reduce the list of candidate variants to further validate.

Data visualisation in IGV

The next step in evaluating the data is to visualise read mappings and variant calls side by side and control that the results make sense for the human eye.

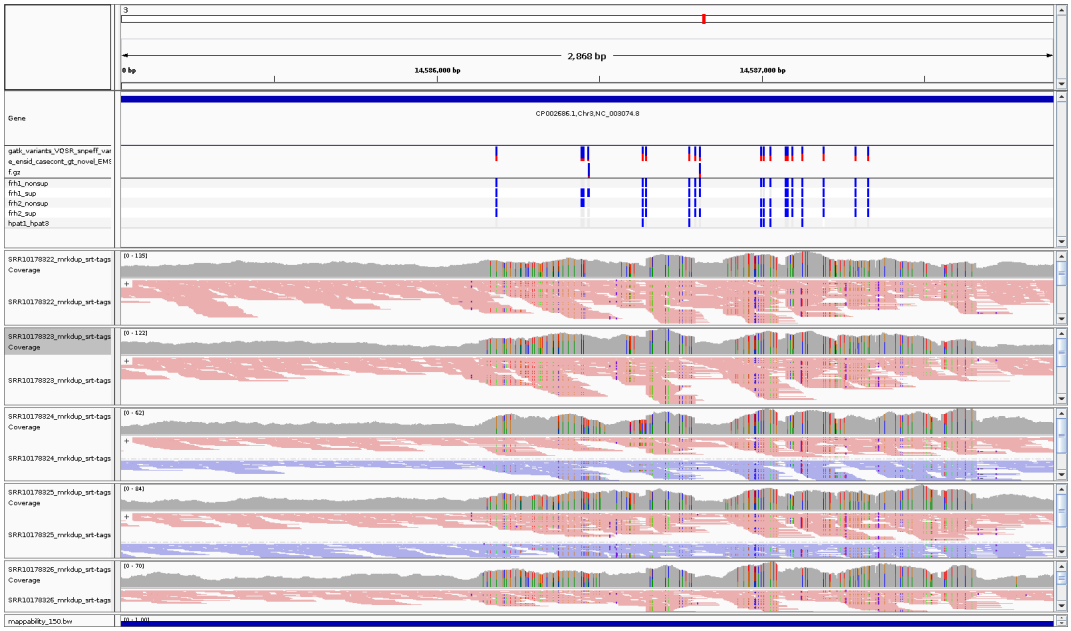
An example of possible sequencing bias is provided next and shows that not all regions of the genome are equal for sequencing due to sequence duplication. Here the Arabidopsis thaliana chloroplast chromosome showing two homologous regions (IRA, IRB) (Sato et al. 1999) in its second half leading to lower mapping quality scores in each block due to the ambiguous mapping.



The mappability track (*Create a mappability track (BITS Wiki)* 2013) shown at the bottom of the figure confirms that short reads of 150b have difficulty finding a unique alignment in these two regions. Such situation often leads to wrong read asignment and calling variants in mis-aligned regions.

FRH1_sup candidates

When exploring the FRH1_sup candidate, it appears that the region around the call is full of other calls.



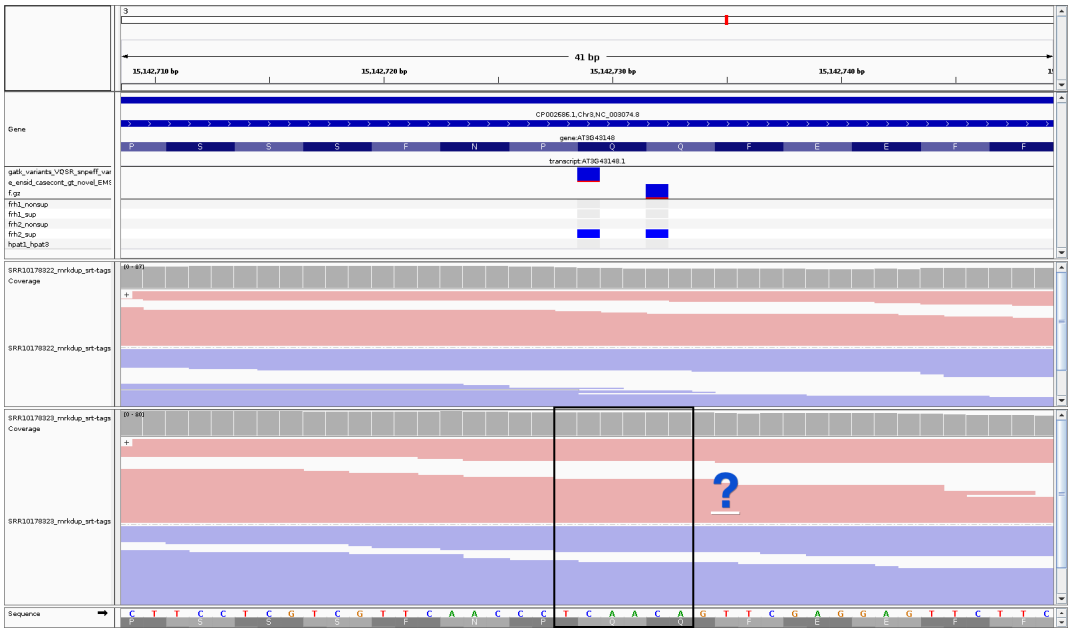
This is a clear evidence for mapping artefact recruting reads to a region where they do not belong. It is likely that this call is not true.

FRH2_sup candidates

The calls for FRH2 candidates are shown next

Legend: The lower track is for the FRH2-sup read mappings, the higher one for the FRH2_nonsup mappings

The first two variants are not confirmed at read level which suggests a software issue during calling.



The other two FRH2-sup specific variants calls are confirmed by the reads mappings.


41 bp
20,935,810 bp
20,935,820 bp
20,935,830 bp
20,935,840 bp
Gene
transcript:AT3G56470.1
gene:AT3G56470
transcript:AT3G08675.1
gene:AT3G08675
gatk_variants...el_EMS.vcf.gz
frh1_nonsup
frh1_sup
frh2_nonsup
SRR10178322_rmlidup_srt-tags
Coverage
SRR10178323_rmlidup_srt-tags
Sequence


41 bp
22,051,040 bp
22,051,050 bp
22,051,060 bp
22,051,070 bp
Gene
IQD13
transcript:AT3G59630.1
transcript:AT3G59630.2
transcript:AT3G59635.1
gene:AT3G59635
gatk_variants...el_EMS.vcf.gz
SRR10178322_rmlidup_srt-tags
Coverage
SRR10178323_rmlidup_srt-tags
Sequence

Conclusion

More candidate will be needed to propose variants for validation in this study, the choice of 'HIGH' impact was possibly too demanding and inspection of less obviously detrimental candidates will be required.

This sequencing approach using pools of F3 plant specimen resulting from a back-cross of phenotyped plants seems a very efficient method to rapidly build candidate lists of reasonable size.

The pooling ensured that minor allele variants and noise were diluted and not represented in the pool in quantifiable amounts while the sequencing of the phenotype-minus population as a separate sample provided additional controls for germline vaiants that may have been introduced before mutagenesis.

This document presents full code and some results of the analysis of 5 public dataset looking for EMS causative variants affecting two different phenotypes. The lack of published results limits the ability to fully evaluate the accuracy of the final results but the bash code provided with this document should ease the analysis of similar experiments

Please report any error or issue to us and suggest improvements so that we can improve ou workflow for future analyses.

last edits: Wed May 06, 2020 v1.0

References

Cingolani, P., A. Platts, l. e. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." *Fly (Austin)* 6 (2): 80–92.

Create a mappability track (BITS Wiki). 2013. VIB BITS. https://wiki.bits.vib.be/index.php/Create_a_mappability_track.

Li, H., and R. Durbin. 2010. "Fast and accurate long-read alignment with Burrows-Wheeler transform." *Bioinformatics* 26 (5): 589–95.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16): 2078–9.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2017. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples," November. Cold Spring Harbor Laboratory. https://doi.org/10.1101/201178.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata. 1999. "Complete structure of the chloroplast genome of Arabidopsis thaliana." *DNA Res.* 6 (5): 283–90.

Thole, J. M., and L. C. Strader. 2015. "Next-generation sequencing as a tool to quickly identify causative EMS-generated mutations." *Plant Signal Behav* 10 (5): e1000167.