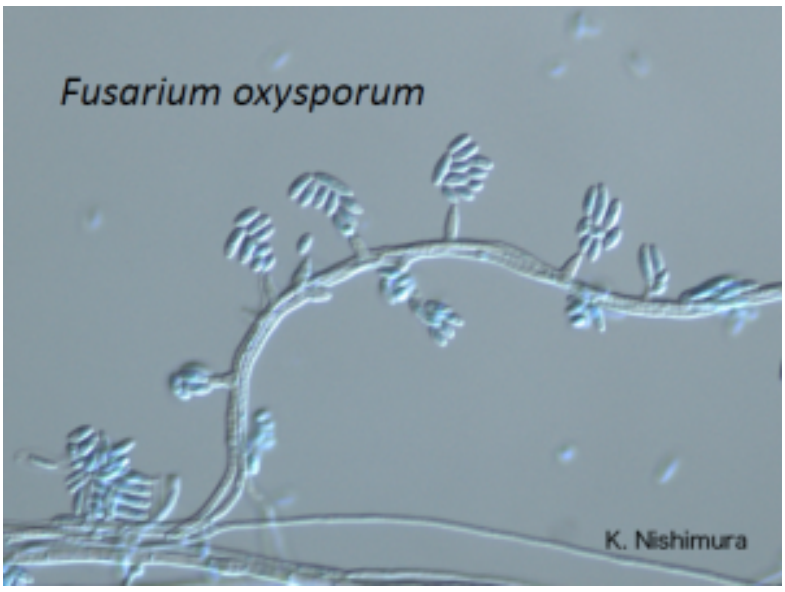
UPPSALA
UNIVERSITET

Identification of *Spok* genes in pathogenic *Fusarium* species

A research training report



Mei Wu

Bioinformatics, MSc
Research training 10 hp, VT2020
Supervisor: Aaron Vogan

Abstract

The genus *Fusarium* encompasses a diverse family of filamentous fungal species in the Ascomycota division. Their prevalence in soil confers inevitable interaction with plants and are subsequently accentuated and brought to attention by the species that are infecting clonal crops. As such, bananas, tomatoes, and others are compromised bringing threat to the agricultural industry. Interestingly, *Fusarium oxysporum*, one of the pathogenic species, infects multiple clonal crops and when various isolates are investigated, they reveal addition of multiple lineage-specific (LS) chromosomes (Ma *et al.* 2010) while existing synteny between 11 core chromosomes. It is unclear whether this is due to horizontal gene transfer or the presence of *Spoks*, which are meiotic drive elements that have been empirically studied in *Podospora anserina* (Vogan *et al.* 2019). It could be other factors driving pathogenicity but until more research is funneled into these fungal genomes or there is more insight on the origins of lineage-specific chromosomes, those are the two hypotheses can be potential explanations. In this report, a combination of methods are explored to find meiotic drive elements, *Spoks*, in *Fusarium oxysporum* population data as doing so would help the scientific world attain a better understanding of their genomes and potential contribution to pathogenicity. My results are a culmination of trial and error, perseverance, and curiosity in what could potentially be driving pathogenicity. I spent 7 weeks with the Systemic Biology department at EBC (Evolutionary Biology Center) to perform such research and in addition to that I learned what it is like to be in a research setting.

Introduction

The research group – Systemic biology

The group, Johannesson's lab, is brought together by an ensemble of people from different backgrounds but all converging onto one ultimate topic with unique contributions. The research group studies many topics relating to filamentous fungi but the topic I am interested in is on meiotic drive elements that are also found in filamentous ascomycetes, specifically, *Neurospora* and *Podospora*. In these sexual eukaryote model systems, meiotic drive elements called Spore killers are found, otherwise known as *Spok* in *Podospora* and Sk-1,2,3 in *Neurospora*. The group aims to use comparative genomics to learn more about these eukaryotes and the like, their sexual cycle, and how selfish elements can act on evolution.

Personnel & weekly activities

I started my 7-week research training course at EBC in the beginning of February with Aaron Vogan, a researcher at Johannesson's lab, as my supervisor. I was to immerse myself in the research environment, learn about the background of the group's research interest, and the project associated with my supervisor. The motivation behind this project stems from his broad interest in the evolution of fungal genomes and how they evolve from one another. He's been with Hannah's group since early 2016 exploring the realm of meiotic drive elements in *Neurospora* and *Podospora* in attempt to understand how these elements impact its population structure. My role in the *Fusarium* project involved working with

various bioinformatic tools/software, building pipelines, and communicating the milestones I have made in the process.

The group held several weekly meetings, all of which I thoroughly enjoyed as they provided an invaluable resource to better understand what was being studied and researched. When it was not an intimate meeting with the group members' giving an update on their project findings, it was learning about transposable elements or a discussion on papers relating to the group's area of interest. The meetings promoted active discussion and learning about the evolving understanding of mating systems, transposable elements, and meiotic drive elements found in fungal genomes.

Common work day

A typical work day involved a lot of reflection and, trial and error with the software in use. It was imperative that I document my progress as this would create an easier reference point for myself and later to communicate the summarization of what I have done in the 7-weeks. If I encountered a problem, it was as easy as messaging my supervisor over Slack, an instant messaging platform for professional use, to gain some clarity on my problem. During downtimes, it was easy to get to know other groups in the department because everyone congregated for lunch and special events that involved food; e.g. national semla day. It is such a well-knit group that even at times of crisis (Covid-19), department members gathered for lunch via webcam and everyone was reachable in person and remotely.

Theory Task

I was given the interesting task of diving deep in the genus similar to *Podospora* and *Neurospora*, *Fusarium* to see if meiotic drive elements are found in this pathogenic genus but due to a limited time frame, I had to narrow my literature and task down to one species, *Fusarium oxysporum*, in which I will refer to as *Fox*. Different clonal crops are at risk of extinction due to specialized strains of *Fox* that are adapted to infect different hosts and the one we're primarily interested in are the ones infecting bananas and tomatoes. They pose as good reference genomes as there exists many studies about them, and they are steadily consumed on a global scale. In a comparative study, it was found that their specialization apart from other species is largely due to an increase of new chromosomes; an additional of four whole chromosomes found in lineage-specific regions of the genome that disrupts the synteny mapped to *F. verticillioides* (a maize pathogen) (Ma et al. 2010). The authors provided three explanations for that and they concluded the most parsimonious explanation is that *Fox* acquired these additional chromosomes through horizontal gene transfer. Interestingly, the *Spok* gene family is hypothesized to undergo horizontal gene transfer as they do not follow the normal fungi phylogeny (Grognet et al. 2014) but until further research is present, this is still up for debate. That said, it would be interesting to see the distribution of *Spok* homologues amongst *Fox* genomes, the location of their distribution, and whether they are along lineage-specific regions or plant effector genes.

Methods

Data acquisition and processing

Spok homologues 1, 3, and 4 from *P. anserina* were obtained from GenBank (Grognet et al. 2014, Vogan et al. 2019) as protein sequences. All of which were queried with BLASTp against a RefSeq Genome database in *Fusarium* (taxid: 5506) to fetch and build a quality

library of *Spok*-like homologues across species. The library of *Fusarium Spoks* would be used to query *Fusarium* population data found in the short-read archive (SRA) (Leinonen et al. 2011) database. I obtained a total of 28 curated sequences in which I carefully selected based on whether long-read sequencing technologies were used or had quality assemblies, and automatically taking in the results that fell below the e-value threshold of 0.05.

Upon retrieval of the desired protein sequences, I translated each of them back to cDNA using NCBI's command line utilities (Kans 2020) by taking each GenBank record's locus tag. The fetched records were in GenBank format and would need further processing to extract the sequence. BioPython's (Cock et al. 2009) SeqIO class was used to extract each sequence and outputted as individual FASTA files. At this point, the reference library is ready to be used as queries against the SRA database.

For my species of interest, I wanted to see the amount of SRA records available so I queried NCBI's taxonomy (Sayers et al. 2009) browser and it shows a total of 1,005 available records that can be iterated through. I would use them against the aforementioned reference library. In order to test if my pipeline worked, each step was performed with scrutiny while querying only one FASTA file against one SRA record. The FASTA file used is the FOXG_14821 locus (https://www.ncbi.nlm.nih.gov/gene/?term=foxg_14821) found in *F. oxysporum* f. sp. *lycopersici* 4287 chromosome 3 and the corresponding SRA record SRR7690004 (https://www.ncbi.nlm.nih.gov/sra/SRX4549585%5Baccn%5D). I used the SRA Toolkit to fetch the record using `fastq-dump` with the option `--split-files` as it is a paired-end genomic file and sequenced using Illumina HiSeq 2500 techonology.

In the reference sequence, I took the flanking regions of 5kb from each end using a python script with options `-e .0001 -f 5000` see (https://github.com/SLAment/Genomics). *Spok* genes are very similar and it is quite possible that there are multiple *Spoks* in the same chromosome or along the genome and therefore, flanking regions are needed to anchor the gene.

Genome alignment and assembly

Next, the reference FASTA sequence was aligned to the SRA record using BWA mem v0.6 (Li 2013) with the options `-P` for paired-end FASTQ file. To allow for faster subsequent processing of the reference, an index file was created using BWA prior to running BWA mem. The output SAM file was piped into samtools v1.6 (Li et al. 2009) using `view` on the SAM files, `sort`, `faidx`, and `index` on the output BAM file. Subsequently, I used `bam2fq` to convert the paired-end sorted bam files to FASTQ but further splitting the output back into two because they will be used as inputs for SPAdes v3.13.0 (Nurk et al. 2013) assembler built for short and long reads.

The modified DNA Illumina reads streamed from samtools was piped into SPAdes to perform de novo assembly. I specified `-s` in front of each paired-end converted sorted bam to fastq file, used the `--careful --cov-cutoff auto` along with k-mers of length `-k 21, 33, 55`.

Results

I used samtools depth to calculate read depth from the aforementioned sorted bam file and it showed a promising coverage at an average of 212.297. Next, I loaded this file in a software known as Integrative Genomics Viewer (IGV) (Robinson et al. 2011) to visualize the

coverage and how it aligns to the reference *Spok*, plus flanking 5 kb sequence (Figure 1). It shows two peaks at opposite ends in a span of 12,936 bp (the size of the reference). I'm not entirely sure how to interpret these results but viewing the coverage track on the screen, I can tell there is a high chance that two *Spoks* exists on the same genome. It seems the variants from positions 5001bp-2936bp (Figure 1.a) are consistently split between two nucleotides. If there was such a possibility then what I mentioned makes sense and BWA had trouble distinguishing the difference between the two because *Spoks* are very similar to one another.

In theory, the *Spok* prior to flanking (size 2,936bp) should be located in the center when viewing on IGV. There would be two peaks of approximately 5kb besides the center and sparse variants along the center. The purpose of using a 5kb flanking region on both ends was to anchor the *Spok* but I did not anticipate that there would be multiple *Spoks* on the genome. Figure 1.b show the peak on the left and Figure 1.c shows the right peak.

In the SPAdes output, only k-mer lengths of 55 had the relevant file (scaffolds.fasta) with consensus sequences but it contained a total of 28,020. This prompted me to BLAST the SPAdes output against a known *Spok* found in this genome using the same aforementioned python script. The script yielded 32 contigs but I took the first longest and meaningful one that showed the presence my query *Spok* in a contig of size 6872. I did the same for the flanked *Spok* reference and I took the top 3 hits. You can see the *Spok* is found adjacent to the 5' end of the flanking region and this is because this was my query for BWA. If it did not capture the other end, it could be that the other flanking region is a transposable element that moved to another part of the genome. Despite this, a *Spok* found further supports my theory that there are two of them present in the genome and SPAdes had trouble distinguishing them apart.

Discussion

Despite that SPAdes could not distinguish the *Spoks* apart found in the contig sequences, I can conclude that this pipeline works in locating robust parts of my query sequence which are the flanking regions that does not experience a lot of contradicting SNPs, and the positive control *Spok* (FOXG_14281) that we expect to find on this genome. The reference *Spok* is found but the view in IGV shows a possibility of another *Spok* in the genome, therefore BWA cannot distinguish which variant belongs to which. I think this would be a matter of tweaking the aligning step with BWA mem. In hindsight, I would be a bit stricter when using the aligner; e.g. increasing the number for mismatch penalties. It could possibly call the correct *Spok* as the aligner becomes less lenient. If the same strain was sequenced using long sequencing technologies became available, I would use that and later correct it with Illumina reads.

If I did not err on that multiple *Spoks* could be found on the same genome and accounting that from the very beginning, I would have more interesting results to look at. Due to a limited time window, I cannot redo the experiment over. However, it would be interesting to see what the distribution of *Spoks* found would look like in a population of *Fox* and later the entire genus's population.

I suspect that if I had insightful results and was able to see a distribution of *Spoks* and specifically amongst pathogenic *Fox* isolates, then I could potentially draw the conclusion that it might have something to do with pathogenicity. If it was found along lineage-specific

regions, then there would be a lot of motivation to study the origins of these regions and whether they have selective outcrossing. Finally, if they sat adjacent to plant-effector genes (Armitage *et al.* 2018) then that is an indication of linkage disequilibrium and being inherited together definitely paints *Spoks* as a potential culprit. It is known that several proteins are secreted (plant-effector) when *F. oxysporum* f. sp. *lycopersici* comes in contact with the tomato plant and two of those pathogens are found on chromosome 14. The entire chromosome 14 is lineage-specific and a location where *Spoks* are found (Ma *et al.* 2010).

All in all, a good understanding of *Spok* distribution would contribution to the existing research regarding these genes and would promote further studies that will potentially tie its association to pathogenicity.

Reflection

My research training practice at EBC with Aaron Vogan was an enjoyable experience. I learned a lot about fungal genomes, what a common workday looks like, and how willingly people are to help one another. All of the weekly meetings were very useful in helping me understand individual projects and their ability to problem solve with dynamic solutions to overcome bottlenecks in their research. They made it really fun to learn about! Communication is imperative in this group and as a newcomer, it worked really well for me. I will definitely take this experience and apply it to my next research training course.

Acknowledgments

I wish to thank Hannah Johannesson for directing me in the right direction for a supervisor, Aaron Vogan for accepting me as the first temporary member of his starting research group and being super awesome, Sandra Lorena Ament for offering her tech-savvy assistance, and other members of Hannah's group that I have not specifically named.

Appendix

qseqid	sseqid	pident	length	qstart	qend	sstart	sstop	evalue
NC_030999.1:935714-938649	NODE_961_length_6872_cov_14.141558	100	848	2089	2936	1	848	0.0
CM000602.2_930714-943649	NODE_961_length_6872_cov_14.141558	100	5848	7089	12936	1	5848	0.0
CM000602.2_930714-943649	NODE_1008_length_6133_cov_10.058736	100	1801	1	1801	4333	6133	0.0
CM000602.2_930714-943649	NODE_2129_length_1569_cov_13.507926	100	1569	2451	4019	1	1569	0.0

Table 1. Query is a *Spok* reference found in *F. oxysporum* f. sp. *lycopersici* 4287 chromosome 14 (FOXG_14281) against the scaffolds produced by SPAdes. The 2nd row is approximately the beginning of the flanking region at the 5' end. The bottom two rows represents the combination of the flanking region at the 3' end.

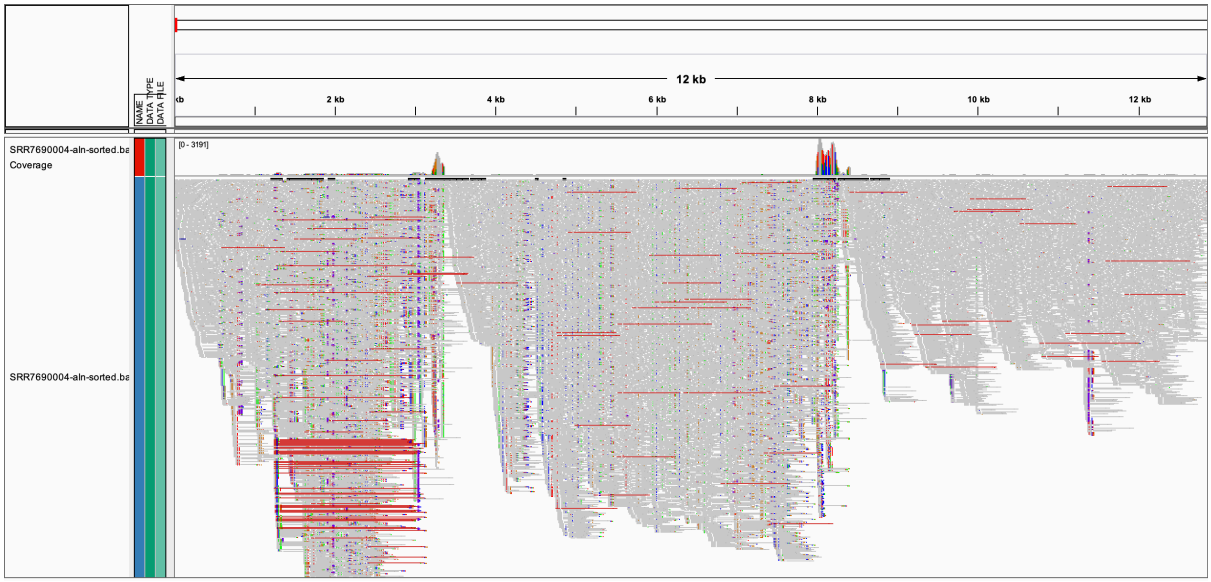


Figure 1.a. A squished and paired view of the flanked reference *Spok* (CM000602.2_930714-943649) aligned to the SRR7690004 record. In between the peaks is where the *Spok* should theoretically lie.

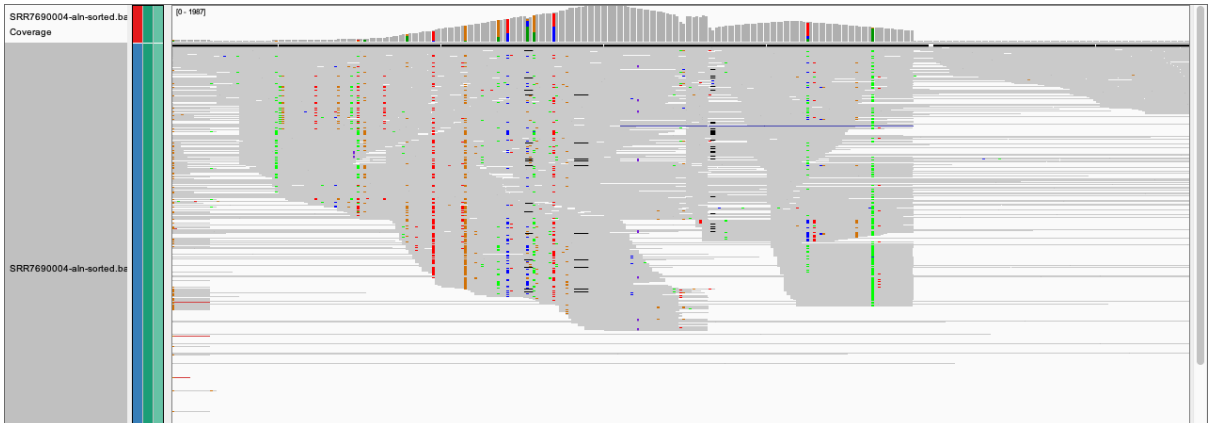


Figure 1.b. A squished and paired view of the left peak of the flanked reference *Spok* (CM000602.2_930714-943649) aligned to the SRR7690004 record.

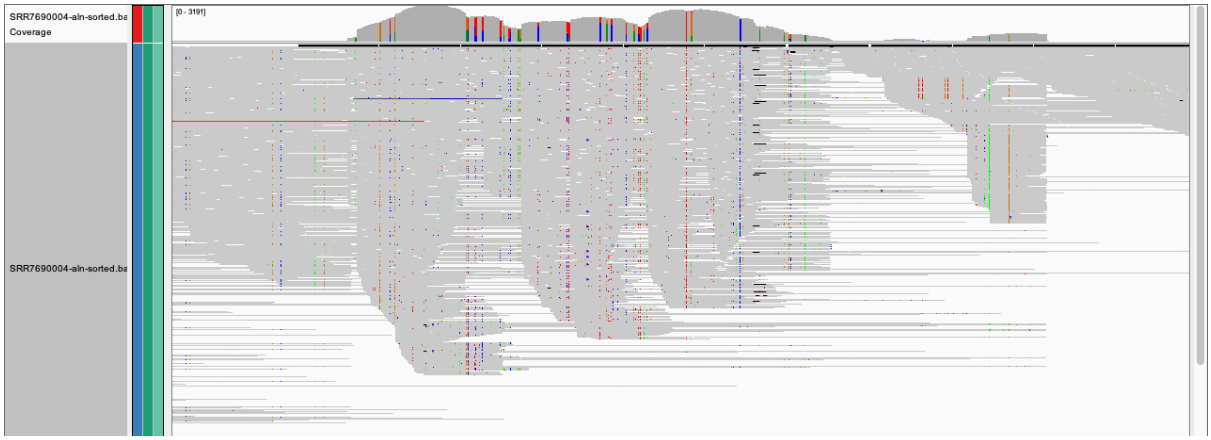


Figure 1.c. A squished and paired view of the right peak of the flanked reference *Spok* (CM000602.2_930714-943649) aligned to the SRR7690004 record.

Literature Cited

Armitage AD, Taylor A, Sobczyk MK, Baxter L, Greenfield BPJ, Bates HJ, Wilson F, Jackson AC, Ott S, Harrison RJ, Clarkson JP. 2018. Characterisation of pathogen-specific regions and novel effector candidates in Fusarium oxysporum f. sp. cepae. Scientific Reports 8: 1–15.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

Grognet P, Lalucque H, Malagnac F, Silar P. 2014. Genes That Bias Mendelian Segregation. PLOS Genetics 10: e1004387.

Kans J. 2020. Entrez Direct: E-utilities on the UNIX Command Line. National Center for Biotechnology Information (US)

Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. Nucleic Acids Research 39: D19–D21.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25: 2078–2079.

Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim W-B, Woloshuk C, Xie X, Xu J-R, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RAE, Chapman S, Coulson R, Coutinho PM, Danchin EGJ, Diener A, Gale LR, Gardiner DM, Goff S, Hammond-Kosack KE, Hilburn K, Hua-Van A, Jonkers W, Kazan K, Kodira CD, Koehrsen M, Kumar L, Lee Y-H, Li L, Manners JM, Miranda-Saavedra D, Mukherjee M, Park G, Park J, Park S-Y, Proctor RH, Regev A, Ruiz-Roldan MC, Sain D, Sakthikumar S, Sykes S, Schwartz DC, Turgeon BG, Wapinski I, Yoder O, Young S, Zeng Q, Zhou S, Galagan J, Cuomo CA, Kistler HC, Rep M. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464: 367–373.

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Deng M, Jiang R, Sun F, Zhang X (ed.). Research in Computational Molecular Biology, pp. 158–170. Springer Berlin Heidelberg, Berlin, Heidelberg.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. Nature biotechnology 29: 24–26.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37: D5-15.

Vogan AA, Ament-Velásquez SL, Granger-Farbos A, Svedberg J, Bastiaans E, Debets AJ, Coustou V, Yvanne H, Clavé C, Saupe SJ, Johannesson H. 2019. Combinations of Spok genes create multiple meiotic drivers in Podospora. eLife 8: e46454.