Additional file 3

This PDF file includes:

1. Supplemental methods

- **(i) Test for recombination**
- **(ii) Tests of functional category enrichment**
- **(iii) Inference of positive selection in *A. fumigatus* virulence genes**

References

Other Additional file for this manuscript includes the following:

Additional file 1

Additional file 2

_____
* Corresponding author
Address correspondence to Prof. Dr. Thomas Dandekar, dandekar@biozentrum.uni-wuerzburg.de

Supplementary methods

Test for recombination

Recombination can inflate the rate of false positives in studies aimed at detecting positive selection based on dN/dS ratios, because the common methods used for detecting positive selection assume that all sites share the same underlying phylogeny [1]. Therefore, we eliminated any genes in the analysis that showed evidence for recombination signals in aligned columns, using a statistical test implemented in GARD program [2] under HyPhy package (Hypothesis testing using Phylogenies) [3]. GARD implements a genetic algorithm for detecting the evidence for potential recombination events from MSA of homologous sequences. Briefly, the method compares a likelihood model assuming a single recombination breakpoint with two different topologies on each side of the breakpoint, with a model that assumes no recombination. If the model with recombination was supported by the Akaike Information Criterion [4], a likelihood ratio test based on the parametric bootstrapping [5] is used to determine whether the model with no recombination can be rejected in favor of a model with recombination. Based on the outcome of the analysis, a level of support is assigned and expressed as a breakpoint placement score [2]. Identified breakpoints by GARD were then assessed for statistical significance by using SH test (Shimodaira-Hasegawa test) [6]. Evidence of significant recombination events were found in 879 alignments (table S15, for details, see Additional file 1 Supplemental Note "Recombination breakpoints"). These clusters where we found statistical evidence of significant breakpoints in alignment, were discarded from further analysis.

Tests of functional category enrichment

GO annotations of *A. fumigatus* single copy ortholog proteins were performed with Blast2GO package [7]. To identify significantly over-represented functional categories, present in the data sets used in this study, Fisher's Exact Test [8] was used. The reference set was constituted of all *A. fumigatus* protein coding genes with GO annotations. To acquire a systematic overview of the PSGs functionalities, we performed the sequence similarity search with BlastP [9] against the KOG database [10]. We used Integrating Network Objects with Hierarchies (INOH) database signal transduction pathway database [11] at InnateDB platform [12] to perform the pathway over-representation analysis of PSGs. DAVID tool [13] was used to calculate whether a certain GO category occurred more often in the PSGs then to be expected looking at all *A.*

fumigatus protein-coding gene annotations at the background, as we did in our recent work [14]. We used level 3 GO terms and a very conservative EASE Score (a modified Fisher Exact p-Value; [15]) threshold 0.05 to identify the significantly over-represented pathways in *A. fumigatus* PSGs. Fisher's exact test was applied to calculate the statistical significance of the over-represented GO categories. The p-value was calculated for Blast2GO, InnateDB and DAVID analysis using hypergeometric tests, and Benjamini-Hochberg adjustment was used for multiple test correction to estimate the proportion of enriched gene sets that would occur by chance given the number of tested gene sets.

Inference of positive selection in *A. fumigatus* virulence genes

A total of 1931 virulence-related genes were collected from literature [16, 17]. A total of 1498 genes in non-SCOs were analysed for positive selection. For reasons of accuracy and computational power, a size reduction approach was followed. In orthogroups with high number of sequences and lower sequence similarity, inference of positive selection is less powerful and less accurate [18] and removal of divergent sequences from the orthogroup for better accuracy can eliminate the high false positive rate caused by highly divergent sequences. Therefore, removal of divergent sequences was performed until the size of 18 sequences was reached. Resulting orthogroups are then aligned first at amino acid level, which are then converted into codon alignments and filtered to remove gap-rich regions. Resulting alignments are then used to infer the phylogenetic gene trees and positive selection with PAML branch-site models.

References

1. Anisimova M, Nielsen R, Yang Z: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164:**1229-1236.
2. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22:**3096-3098.
3. Pond SL, Frost SD, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21:**676-679.
4. Burnham KP, Anderson DR, Huyvaert KP: **AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons.** *Behavioral Ecology and Sociobiology* 2010, **65:**23-35.
5. Goldman N: **Statistical tests of models of DNA substitution.** *J Mol Evol* 1993, **36:**182-198.
6. Shimodaira H, Hasegawa M: **Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference.** *Molecular Biology and Evolution* 1999, **16:**1114-1116.

7. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
8. Connelly LM: **Fisher's Exact Test.** *Medsurg Nurs* 2016, **25:**58, 61.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
10. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28:**33-36.
11. Yamamoto S, Sakai N, Nakamura H, Fukagawa H, Fukuda K, Takagi T: **INOH: ontology-based highly structured database of signal transduction pathways.** *Database (Oxford)* 2011, **2011:**bar052.
12. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock RE, Brinkman FS, Lynn DJ: **InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation.** *Nucleic Acids Res* 2013, **41:**D1228-1233.
13. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
14. Srivastava M, Bencurova E, Gupta SK, Weiss E, Loffler J, Dandekar T: **Aspergillus fumigatus Challenged by Human Dendritic Cells: Metabolic and Regulatory Pathway Responses Testify a Tight Battle.** *Front Cell Infect Microbiol* 2019, **9:**168.
15. Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4:**R70.
16. Puertolas-Balint F, Rossen JWA, Oliveira Dos Santos C, Chlebowicz MMA, Raangs EC, van Putten ML, Sola-Campoy PJ, Han L, Schmidt M, Garcia-Cobos S: **Revealing the Virulence Potential of Clinical and Environmental Aspergillus fumigatus Isolates Using Whole-Genome Sequencing.** *Front Microbiol* 2019, **10:**1970.
17. Vivek-Ananth RP, Mohanraj K, Vandanashree M, Jhingran A, Craig JP, Samal A: **Comparative systems analysis of the secretome of the opportunistic pathogen Aspergillus fumigatus and other Aspergillus species.** *Sci Rep* 2018, **8:**6617.
18. Wong WS, Yang Z, Goldman N, Nielsen R: **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites.** *Genetics* 2004, **168:**1041-1051.