bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Perturbed human sub-networks by *Fusobacterium nucleatum* candidate virulence factors

Andreas Zanzoni[1,#], Shérazade Braham[1], Lionel Spinelli[1], Christine Brun[1,2]

[1]Aix-Marseille Université, Inserm, TAGC UMR_S1090, Marseille, France;

[2]CNRS, Marseille, France.

Running Head: The inferred *Fuso*Secretome – human interaction network.

Address correspondence to A. Zanzoni, andreas.zanzoni@inserm.fr

Keywords: Fusobacterium nucleatum, secretome, molecular mimicry, short linear motifs, bioinformatics, interaction network, colorectal cancer, inflammatory bowel's diseases, virulence factors.

Abstract word count: 207 words.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

F. nucleatum is a gram-negative anaerobic species residing in the oral cavity and implicated in several inflammatory processes in the human body. Although *F. nucleatum* abundance is increased in inflammatory bowel's disease subjects and is prevalent in colorectal cancer patients, the causal role of the bacterium in gastrointestinal disorders and the mechanistic details of host cell functions subversion are not fully understood.

We devised a computational strategy to identify putative secreted *F. nucleatum* proteins (*Fuso*Secretome) and to infer their interactions with human proteins based on the presence of host molecular mimicry elements. *Fuso*Secretome proteins share similar features with known bacterial virulence factors thereby highlighting their pathogenic potential. We show that they interact with human proteins that participate in infection-related cellular processes and localize in established cellular districts of the host-pathogen interface. Our network-based analysis identified 31 functional modules in the human interactome preferentially targeted by 138 *Fuso*Secretome proteins, among which we selected 26 as candidate virulence factors, representing both putative and known virulence proteins. Finally, 6 of preferentially targeted functional modules are implicated in the onset and progression of inflammatory bowel's diseases and colorectal cancer.

Overall, our computational analysis identified candidate virulence factors potentially involved in the *F. nucleatum* – human cross-talk in the context of gastrointestinal diseases.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Fusobacterium nucleatum is a gram-negative anaerobic bacterium best known as a component of the oral plaque and a key pathogen in gingivitis and periodontitis [1]. It has also been isolated in several inflammatory processes in distinct body sites (*e.g.*, endocarditis, septic arthritis, liver and brain abscesses) and implicated in adverse pregnancy outcomes (reviewed in [2]). Moreover, it has been demonstrated that *F. nucleatum* can adhere to and invade a variety of cell types, thereby inducing a pro-inflammatory response [3–8]. Recent works showed that *(i) F. nucleatum* is prevalent in colorectal cancer (CRC) patients [9–11] and *(ii)* its abundance is increased in new-onset Crohn's disease (CD) subjects [12]. Interestingly, following studies suggested a potential role of this bacterium in CRC tumorigenesis and tumor-immune evasion [13–16].

Despite these findings, a large fraction of *F. nucleatum* gene products are still uncharacterized. Moreover, to date, only a handful of pathogenic factors have been experimentally identified [17,18] and protein interaction data between these factors and human proteins, which could inform on the molecular details underlying host-cell subversion mechanisms, are sparse [4,16,19]. Altogether, this underlines that a comprehensive view of the molecular details of the *F. nucleatum* – human cross-talk is currently missing.

How *F. nucleatum* could hijack human cells? Pathogens employ a variety of molecular strategies to reach an advantageous niche for survival. One of them consists in subverting host protein interaction networks. Indeed, they secrete and deliver factors such as toxic compounds, small peptides, and even proteins to target the host molecular networks. To achieve this, virulence factors often display structures resembling host components in form and function [20–22] to interact with host proteins, thus providing a benefit to the pathogen [23]. Such "molecular mimics" (*e.g.*, targeting motifs, enzymatic activities and protein-protein interaction elements) allow pathogens to enter the host cell and perturb cell pathways (*e.g.*, [24–26]).

Over the years, several experimental approaches have been applied to identify protein-protein interactions (PPIs) between pathogen and their hosts providing

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

new insights on the pathogen's molecular invasion strategies. However, the vast majority of these systematic studies focused on viruses (e.g., [27–29]) and, to a lesser extent, on bacteria [30–33] and eukaryotic parasites [33,34]. Indeed, as cellular pathogens have large genomes and complex life cycles, the experimental identification of virulence factors and the large-scale mapping of host-pathogen PPIs require a lot of effort and time [35,36]. In this context, computational approaches have proved to be instrumental for the identification of putative pathogenic factors (*e.g.*, [37,38]), the characterization of molecular mimics [23,39,40] and the inference of their interactions with host proteins (for a review see [41]).

Here, in order to gain new insights on the molecular crosstalk between the *F. nucleatum* and the human host, we devised a computational strategy combining secretion prediction, protein-protein interaction inference and protein interaction network analyses (Figure 1). Doing so, we defined a secretome of the bacterium and the human proteins with whom they interact based on the presence of mimicry elements. We identified the host cellular pathways that are likely perturbed by *F. nucleatum* including immune and infection response, homeostasis, cytoskeleton organization and gene expression regulation. Interestingly, our results identify candidate virulence factors, including the established Fap2 adhesin, and provide new insights underlying the putative causative role of *F. nucleatum* in colorectal cancer and inflammatory bowel's diseases.

Results

Prediction of *F. nucleatum* secreted proteome. Previous computational analyses highlighted that *F. nucleatum* has a reduced repertoire of secretion machines [42,43] meaning that it might exploit alternative "non-classical" translocation mechanisms to unleashed virulence factors. Thus, we sought to identify putative *F. nucleatum* secreted proteins by analyzing the 2046 sequences of *F. nucleatum subsp. nucleatum* (strain ATCC 25586) proteome

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

using two distinct algorithms: SignalP [44] for peptide-triggered secretion and SecretomeP [45] for leaderless protein secretion. While the SignalP algorithm predicted 61 *F. nucleatum* sequences being secreted *via* classical/regular secretion pathways, SecretomeP found 176 proteins as possibly secreted through non-classical routes. In total, we identified 237 putative secreted proteins in the *F. nucleatum* proteome (herein called "*Fuso*Secretome") (Table S1). Notably, we were able to correctly predict as secreted all the *F. nucleatum* virulence factors known so far, namely FadA (FN0264), Fap2 (FN1449), RadD (FN1526) and the recently identified Aid1 adhesin (FN1253) [46]. This result underlines the relevance of secretion prediction to identify novel putative virulence factors in the *F. nucleatum* proteome.

It has been shown that disorder propensity is an emerging hallmark of pathogenicity [47,48]. By analyzing the intrinsic disorder content of the *Fuso*Secretome proteins, we indeed observed a significantly higher disorder propensity of these proteins compared to the non-secreted proteins (P-value = 1.9×10^{-4}, Kolgomorov-Smirnov test, two-sided) (Figure 2; Figure S1), further reinforcing their possible role in the infection/invasion process.

To detect functional elements that can further contribute to *F. nucleatum* pathogenicity, we sought for the presence of globular domains in the *Fuso*Secretome. We observed an enrichment of domains mainly belonging to the outer membrane beta-barrel protein superfamily (Table 1). Six out of the eight over-represented domains among the the *Fuso*Secretome proteins are also found in known virulence factors of gram-negative bacteria [49] and are involved in adhesion, secretion, transport and invasion. Altogether, these findings suggest that *Fuso*Secretome proteins display features of known virulence factors and can likely be involved in the cross-talk with the human host.

Inference of the *Fuso*Secretome – human interaction network. Generally, pathogens employ a variety of molecular stratagems to interfere with host-cell networks, controlling key functions such as plasma membrane and cytoskeleton dynamics, immune response and cell death/survival. In particular, their proteins

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

often carry a range of *mimics*, which resemble structures of the host at the molecular level, to 'sneak' into host cells [20–22,50].

Here, we focused on putative molecular mimicry events that can mediate the interaction with host proteins: *(i)* globular domains that occur in both *Fuso*Secretome and the human proteome and *(ii)* known eukaryotic short linear motifs (SLiMs) found in *Fuso*Secretome proteins. SLiMs are short stretches of 3-10 contiguous amino acids residues that often mediate transient PPIs and tend to bind with low affinity [51]. We scanned the sequences of the *Fuso*Secretome and human proteins for the presence of domains as defined by Pfam [52], and detected the occurrence of experimentally identified SLiMs gathered from the Eukaryotic Linear Motif (ELM) database [53] (see Methods).

Doing so, we identified 55 "human" domains in 50 *Fuso*Secretome proteins out of 237, including several domains related to ribosomal proteins, aminopeptidases and tetratricopeptide repeats (TPR) (Table S3). Interestingly, 29 of these domains are also found in known bacterial binders of human proteins [30]. Furthermore, 139 *Fuso*Secretome proteins contain at least one instance of SLiMs. Most of the 57 different detected SLiMs represents binding sites such as motifs recognized by PDZ, SH3 and SH2 domains (Table S3).

We exploited these putative mimicry events to infer the interaction with human proteins by using templates of domain-domain and SLiM-domain interactions (see Methods for further details). Doing so, we obtained 3,744 interactions (1,544 domain- and 2,201 SLiM-mediated interactions, respectively) between 144 *Fuso*Secretome and 934 human proteins (Table S4 and Table S5).

In order to assess the reliability of the inferences, we evaluated the biological relevance of the putative human interactors using different orthogonal datasets. First, human proteins experimentally identified as binders or targets of bacterial and/or viral proteins are over-represented among the 934 inferred human interactors of the *Fuso*Secretome proteins (415 proteins, 1.3-fold, P-value=1.61×10^{-11}). Notably, the over-representation holds when bacterial and viral binders are considered separately (176 bacterial interactors, 1.1-fold, P-value = 3.5×10^{-3} and 338 viral interactors, 1.5 enrichment ratio, P-value< $2.2 \times 10^{-}$

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[16]). Second, according to the Human Proteins Atlas (see Methods), the majority of the inferred human interactors has been detected either in the saliva (673, 72%), small intestine (652, 70%) or colorectal (671 proteins, 72%) tissues, confirming their presence in human tissues hosting *F. nucleatum.* Third, we assessed whether the inferred human interactors are implicated in gastrointestinal disorders by seeking for an over-representation of genes associated to such diseases (see Methods). Indeed, the human interactors of the *Fuso*Secretome are enriched in *(i)* proteins identified in the human colon secretomes of colorectal cancer (CRC) tissue samples (3.5-fold, P-value < 2.2×10^{-16}), *(ii)* proteins encoded by genes whose expression correlates with *F. nucleatum* abundance in CRC patients (2-fold, P-value=4×10^{-4}) and *(iii)* genes associated with inflammatory bowel's diseases (IBDs) (2-fold, P-value=8×10^{-4}). Altogether, the results of these analyses highlight the relevance of the inferred human interactors as putative binders of *F. nucleatum* virulence factors and their potential implication in gut diseases, therefore validating the undertaken inference approach.

Functional role of the human proteins targeted by *F. nucleatum*. Globally, the inferred *Fuso*Secretome human interactors are involved in several processes related to pathogen infection such as immune response and inflammation, response to stress, endocytosis as shown by the 137 significantly enriched Biological Processes Gene Ontology (GO) terms among their annotations (Table 2, Table S6). Similarly, the targeted human proteins are over-represented in 125 pathways (54 from KEGG and 71 from Reactome databases [54,55]) involved in cell adhesion and signaling, extracellular matrix remodeling, immunity, response to infection and cancer-related pathways (Table 2, Table S6). These human proteins are mainly localized in the extracellular space, plasma membrane and at cell-cell junctions that represent the main districts involved in the initial encounter between a pathogen and the host, as indicated by the over-representation of 30 Cellular Component GO terms (Table 2, Table S6). Overall, this indicates that our inferred interactions can participate in the *F.nucleatum* – human crosstalk.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

***F. nucleatum* targets topologically important proteins in the host network.** To gain a broader picture of the inferred interactions in the cellular context, we mapped the *Fuso*Secretome human interactors on a human binary interactome that we previously built from publicly available protein-protein interaction data [56] (Table S7). Around 70% of the inferred human interactors (*i.e.*, 663 proteins) are present in the human binary interactome. Interestingly, the human targeted proteins occupy topologically important position in the interactome as shown by their significantly higher number of interactions and higher values of betweenness centrality compared to other network proteins (number of interactions: mean=23 vs. 11, P-value=$1.9x10^{-10}$; betweeness centrality: mean=0.00078 vs. 0.00018, P-value=$6.2x10^{-12}$) (Figure 3).
The human interactome is composed of functional network modules, defined as group of proteins densely connected through their interactions and involved in the same biological process [57] (see Methods). We thus next investigated the 855 functional modules that we previously detected [56] using the OCG algorithm [58] (Table S8). A significant number of interactors participate in 2 or more of these functional units (259 proteins, 1.3-fold enrichment, P-value=$1.4x10^{-7}$), indicating that the *Fuso*Secretome tends to target multifunctional proteins in the human interactome [58]. Moreover, among the multifunctional inferred human interactors we found an enrichment of extreme multifunctional proteins (52 interactors, 2-fold enrichment, P-value=$1.0x10^{-5}$), which are defined as proteins involved in unrelated cellular functions and may represent candidate moonlighting proteins [59]. This suggests that the *Fuso*Secretome might perturb multiple cellular pathways simultaneously by targeting preferentially a whole range of multifunctional proteins.

Functional subnetworks of the human interactome perturbed by *F. nucleatum* and identification of the main virulence factors. Based on their enrichment in inferred human interactors, 31 network modules (~4% of the 855 detected modules) are preferentially targeted by 138 distinct proteins of the

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

*Fuso*Secretome (Table 3). Targeted modules are involved in relevant processes such as immune response, cytoskeleton organization, cancer and infection-related pathways (Table 3 and Table S9). Moreover, proteins belonging to these modules are mainly localized in the extracellular space or in membranous structures (Table 3 and Table S9), which represent important districts of the microbe-host interface. Interestingly, the enrichment of functional categories related to gene expression regulation (Table S9) in several modules suggests novel potential host subversion mechanisms by *F. nucleatum*.

These modules are targeted on average by 50 FusoSecretome proteins (ranging from 2 to 104 per module) and the number of inferred host-pathogen interactions for each module varies considerably (Table 3). What are the main perturbation factors among the *Fuso*Secretome proteins? To quantify their impact on modules based on the number of interactions they have with each of them, we computed a Z-score (see Methods). We considered the 26 *Fuso*Secretome proteins having a perturbation Z-score > 2 in at least one module as candidate virulence factors. They consist in outer membrane proteins, enzymes, iron-binding proteins and protein involved in transport (Table 4). Ten of them (38%) can perturb at least two distinct modules (Figure 4A). Notably, we identified among the candidates, the known virulence factor Fap2 (FN1449) (Figure 4B) that targets 4 modules, and a protein containing the MORN_2 domain (FN2118) (Figure 4C) recently identified as a key element in actively invading *F. nucleatum* species [60], which perturbs 6 modules. On the other hand, 25 preferentially targeted modules are perturbed by at least two candidate virulence factors, Module 78 involved in immune response being the most potentially subverted (Figure 4A).

***F. nucleatum* and gut diseases from a network perspective.** Among the 855 network modules detected in the human interactome, 38 are enriched in genes involved in at least one gut disease (*i.e.*, CRC and IBDs, see Methods). Interestingly, 27 of them (*i.e.*, 71%) are targeted by at least one *Fuso*Secretome protein, among which 3 contain a statistically significant fraction of inferred

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

human interactors (Figure 5). Notably, Module 78, involved in immune response, is enriched in genes associated to inflammatory bowel's diseases (IBDs) (28 proteins, 5.2-fold enrichment, P-value= $4.78x10^{-4}$) as well as in CD-specific (9 proteins, 13.4-fold enrichment, P-value=$2.52x10^{-3}$) and CRC-mutated (11 proteins, 4-fold enrichment, P-value=$1.46x10^{-2}$) genes. Moreover, it is enriched in genes whose expression correlates with *F. nucleatum* abundance in CRC patients (24 proteins, 3.7-fold enrichment, P-value=$3.35x10^{-4}$). This module is targeted by several candidate virulence factors, including an hemolysin (FN0291), an outer membrane protein (FN1554) and the MORN_2 domain containing protein (FN21118) (Figure 4A), which therefore, may play critical roles in these diseases. IBD genes are also enriched in Module 702 (5 proteins, 11-fold enrichment, P-value=$2.13x10^{-2}$), whose proteins participate in Jak-STAT signaling, whereas CD-specific genes are over-represented in Module 9 implicated in immunity (5 proteins, 28-fold enrichment, P-value=$2.52x10^{-3}$). Interestingly, Module 9 is specifically perturbed by Fap2 (FN1449) (Figure 4B), which is known to modulate the host immune response.

Three other modules enriched in inferred human interactors show a significant dysregulation of the expression of their constituent proteins during CRC progression [61] and are implicated in infection-response pathways and cytoskeleton organization (Figure 5). In particular, two of these modules (Modules 138 and 216) show significant and specific up-regulation in Stage II, whereas the third (Module 371) is significantly up-regulated in normal and Stage II samples. Overall, these results indicate that *F. nucleatum* could contribute to the onset and progression of IBDs and CRC by perturbing some of the underlying network modules.

Discussion

Over the years, it has been shown that *F. nucleatum* can adhere and invade human cells triggering a pro-inflammatory response. Nevertheless, the current knowledge on the molecular players underlying the *F. nucleatum* – human cross-talk is still limited.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For this reason, we carried out a computational study to identify *F. nucleatum* putative secreted factors (*Fuso*Secretome) that can interact with human proteins. The originality of our study is manifold compared to previous work. First, we used secretion prediction to identify potential *F. nucleatum* proteins that can be present at the pathogen-host interface. Second, we exploited both domain-domain and domain-motif templates to infer protein interactions with human proteins. Earlier works, including one on *F. nucleatum*, chiefly applied homology-based methods for host-pathogen interaction inference (*e.g.*, [62–65]). To our knowledge, domain-motif templates has been only exploited so far to infer or to resolve human-virus protein interaction networks [39,66]. Indeed, SLiM mimicry is widespread among viruses [21,67], but increasing evidence shows that it can be an effective subversion strategy in bacteria as well [22]. Third, we performed a network-based analysis on the human interactome to identify candidate *F. nucleatum* virulence factors and the sub-networks they likely perturb.

Our approach relies on two prediction steps: *(i)* the definition of the *Fuso*Secretome based either on the presence of a signal peptide or several protein features such as disorder content, and *(ii)* the detection of host mimicry elements involved in the interaction with the host. It could be argued that the SecretomeP algorithm may incorrectly predict some proteins as secreted because of their high disorder content. For instance, a previous study considered as erroneous the secretion prediction of ribosomal proteins [68]. We assigned 20 ribosomal proteins to the *Fuso*Secretome. Although we cannot exclude a misprediction, ribosomal proteins can be secreted in some bacteria and involved in host interaction [69,70]. Furthermore, increasing evidence shows that ribosomal proteins are moonlighting proteins with extra-ribosomal functions such as the *E. coli* ribosomal L2 protein that moonlights by affecting the activity of replication proteins [71]. Among the 337 inferred interactions between the 20 *Fuso*Secretome ribosomal and 183 human proteins, only a third of them belongs to ribosomal protein families. Interestingly, only 3 of the 41 human interactors inferred for *F. nucleatum* L2 are ribosomal proteins, and we identified the L2

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein as candidate effector preferentially targeting Module 451. As this module is mainly involved in cell cycle and DNA repair, this result evokes the ability of L2 in *E. coli* to interfere with DNA processing factors [71] and further reinforces the confidence in the secretome prediction. Moreover, we have here underlined the value of the proposed approach: the interactome provides, on the one hand, the proper biological context to filter out potential false positive inferred interactions and, on the other, pinpoints candidate proteins that can be involved in the *F. nucleatum* – host interface.

Concerning the host mimicry elements, SLiM detection is notorious for over-prediction [72], given their relative short length and degeneracy (*i.e.*, few fixed amino acid positions). Our strategy to control for false positives was to consider only SLiM occurrences in the *Fuso*Secretome protein regions predicted as disordered. Indeed, the vast majority of known functional SLiMs falls in unstructured regions [72,73]. Conversely, we might also have missed some "true" mimicry instances in the *Fuso*Secretome by using too stringent parameters for domains and SLiMs identification and our predictions may well be incomplete. However, their functional significance fortifies our confidence in the predictive approach. Indeed, the *Fuso*Secretome shares similar features with known virulence factors highlighting its pathogenic potential. In addition, interactors are implicated in established biological processes and cellular districts of the host-pathogen interface and significantly overlap with known pathogen protein binders. Furthermore, more than 70% of interactors are expressed in either the saliva or intestinal tissues. This suggests that the vast majority of the inferred interactions can occur in known *F. nucleatum* whereabouts in the human body. Finally, we found among the human interactors an over-representation of genes whose expression correlates with *F. nucleatum* in CRC patients [13] as well as in IBD-related genes [74], which are mainly involved in immune- and infection-response pathways.

Moreover, we gained a more comprehensive view of the cellular functions that can be perturbed by the *Fuso*Secretome by investigating the human interactome. In agreement with previous experimental observations of host cell networks

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

targeted by distinct pathogens, *F. nucleatum* targets hubs and bottlenecks in the human interactome [30,33,75]. Interestingly, *Fuso*Secretome tend to interact with multifunctional proteins. This can represent an effective strategy to interfere with distinct cellular pathways as the same time [76].

Among the network modules preferentially targeted by the *Fuso*Secretome, we identified, besides the well-established functions related to host – pathogen interactions, several modules involved in chromatin modification and transcription regulation (Modules 246, 451, 571 and 625), and localized in compartments such as perinuclear region of the cytoplasm (Modules 90, 138 and 615). Intriguingly, this is reminiscent of the fact that invading *F. nucleatum* strains localize in perinuclear district of colorectal adenocarcinoma cells [8], and that bacteria can tune host-cell response by interfering directly – or indirectly – with the chromatin organization and the regulation of gene expression [77].
We propose 26 *Fuso*Secretome proteins as candidate virulence factors. They are the predominant interactors of preferentially targeted modules. Among the candidates we identified the known virulence factor Fap2, which was recently shown to promote immune system evasion by interacting with the immunoreceptor TIGIT [19]. Interestingly, Fap2 interacts specifically with Module 9, which is involved in immune response, thus suggesting novel potential binders mediating Fap2 subversion.

A recent report found that abundance of *F. nucleatum* is associated with high microsatellite instability tumors and shorter survival [14]. Notably, three preferentially targeted network modules (*i.e.*, Modules 138, 216 and 371) show a significant up-regulation in a stage associated to high microsatellite instability during CRC progression (stage II) [61] [78] and poor prognosis [79] [80]. This suggests that these modules may be important for CRC progression and outcome, and that the inferred interactions targeting these modules can mediate the cross-talk between *F. nucleatum* and the host in this particular subtype of CRC.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Overall, our functional and network-based analysis shows that the proposed interactions can occur *in vivo* and be biologically relevant for the *F. nucleatum* – human host dialogue.

Conclusions

Over the last years, many microbes have been identified as key players in chronic disease onset and progression. However, untangling these complex microbe-disease associations requires lot effort and time, especially in the case of emerging pathogens that are often difficult to manipulate genetically. By detecting the presence of host mimicry elements, we have inferred the protein interactions between the putative secretome of *F. nucleatum* and human proteins, and ultimately provided a list of potential virulence factors that can be involved in the *F. nucleatum* – human cross-talk. Our computational strategy can be helpful in guiding and speeding-up wet lab research in microbe-host interactions.

Methods

Protein sequence data. The reference proteomes of *Fusobacterium nucleatum subsp. nucleatum* strain ATCC 25586 (Proteome ID: UP000002521) and *Homo sapiens* (Proteome ID: UP000005640) were downloaded from the UniProtKB proteomes portal [81] (April 2013). The protein sequences of known gram-negative bacteria virulence factors were taken from the Virulence Factors DataBase [49] (January 2014).

Secretome prediction. We identified putative secreted proteins among the *F. nucleatum* proteins by applying two algorithms: SignalP 4.1 [44] that detects the presence of a signal peptide and SecretomeP 2.0 [45] that identifies non-classical secreted proteins (i.e. not triggered by a signal peptide) using a set of protein features such as amino acid composition and intrinsic disorder content.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Disorder propensity. To evaluate the intrinsic disorder propensity of *F. nucleatum* proteins predicted as secreted, we used the stand-alone programs of the following algorithms: DISOPRED (version 2.0) [82], IUPred (both long and short predictions) [83] and DisEMBL (COILS and HOTLOOPS predictions, version 1.4) [84]. We compared the disorder propensity distribution of SignalP-predicted secreted proteins to non-secreted proteins using the Kolmogorov-Smirnov test (two-sided, alpha=0.05).

Detection of functional domains. We ran the pfamscan program [85] on *F.nucleatum*, *H. sapiens* and virulence factors protein sequences to detect the presence of Pfam domains [52] (release 26). We kept only Pfam-A matches with an E-value<10^{-5}.

Identification of short linear motifs. We used the SLiMSearch 2.0 tool [86] to identify occurrences of known short linear motifs from the ELM database [53] (downloaded in May 2013) in the *F. nucleatum* proteome. As functional ELM motifs tend to be located in disordered regions [87], we applied the protein disorder context filter in SLIMSearch (average motif disorder score > 0.2) to select putative relevant motif occurrences.

Protein interaction inference. We built an interaction network between *F. nucleatum* putative secretome and human proteins by using interaction templates from the 3did database [88], which stores high-resolution three-dimensional templates for domain-domain interactions, and the iELM resource [89] that lists motif-mediated interactions respectively. Both datasets were downloaded in August 2013. The domain-based interaction inference works as follow: given a pair of known interacting domains *A* and *B*, if domain *A* is detected in the *F. nucleatum* protein *a* and domain *B* in the human protein *b*, then an interaction between *a* and *b* is inferred. Analogously, for the SLiM-mediated interaction inference: for a given known ELM motif *m* interacting with the domain *C* in the

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

human protein *c*, if the motif *m* occurs in the *F. nucleatum* protein *a*, then *a* is inferred to interact with *c*.

Human proteins targeted by bacteria and viruses. We gathered a list of 3,428 human proteins that were experimentally identified as interaction partners of three bacterial pathogen proteins (*Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*) in a large-scale yeast two-hybrid screen [30]. We downloaded interaction data with viruses for 4,897 human proteins from the VirHostNet database [90].

Functional enrichment analysis. We performed enrichment analysis on the list of inferred human interactors using the g:Profiler webserver [91] (build date: July 2013). We analyzed the following annotations: Biological Process and Cellular Component from the Gene Ontology [92]; biological pathways from KEGG [54] and Reactome [55]. Functional categories containing less than 5 and more than 500 genes where discarded. P-values were corrected for multiple testing using the Benjamini-Hochberg procedure (significance threshold = 0.025).

Human expression data. RNA-seq expression data for 20,345 protein coding genes in normal colorectal, salivary gland and small intestine (i.e., jejunum and ileum) tissues was downloaded from the Human Protein Atlas (version 13), a compendium of gene and protein expression profiles in 32 tissues [93]. We considered as expressed those protein-coding genes with a FPKM>1, that is 13,640 for colorectal, 13,742 for salivary gland and 13,220 for small intestine. Patient secretome profiling (2,566 proteins) for tumor colorectal tissue samples were taken from [94].

Gut-related disease gene sets. We retrieve 152 colorectal cancer genes from the Network of Cancer Genes database (version 4.0, [95]). The list of human genes whose expression correlates with *F. nucleatum* abundance in colorectal cancer patients [13] was kindly provided by Aleksandar Kostic (Broad Institute,

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

USA). The compendium of 163 loci associated with inflammatory bowel diseases was taken from a large meta-analysis of Crohn's disease and ulcerative colitis genome-wide association studies [74].

Human interactome building, network module detection and annotation. We use the human interactome that we assembled and used in [56,61]. Briefly, protein interaction data were gathered from several databases through the PSICQUIC query interface [96]. We kept only direct (*i.e.,* binary) interactions according to the experimental detection method, and reduced sequence redundancy at 95% identity among the proteins in the interactome using the CD-HIT algorithm [97]. In this way, we obtained a human binary interactome containing 74,388 interactions between 12,865 proteins (February 2013).
A collection of 855 network modules detected using the Overlapping Cluster Generator algorithm [58] was functionally annotated by assessing the enrichment of Gene Ontology (GO) biological process and cellular component terms [92], and cellular pathways from KEGG [54] and Reactome [55]. Enrichment P-values were computed using the R package gProfileR [91] and corrected for multiple testing with the Benjamini–Hochberg procedure. Similarly, the over-representation of inferred human interactors and gut disease gene sets in network modules of the human interactome was assessed using a one-sided Fisher's exact test followed by Benjamini–Hochberg multiple testing correction (significance threshold = 0.025).

Network module perturbation Z-score. We devised a score to quantify the contribution of *F. nucleatum* secreted proteins to the perturbation of a network module through their inferred interactions. We defined the perturbation Z-score for each *F. nucleatum* protein *f* interacting with at least one protein in module *m* as follows:

$$Z_{f,m} = \frac{x_{f,m} - \mu_m}{\sigma_m}$$

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Where $x_{f,m}$ is the number of inferred interactions of the protein *f* with module *m*, $Z_{f,m}$ is the perturbation Z-score of the protein *f* in the module *m*, μ_m and σ_m are the mean of the inferred interaction values and their standard deviation in the module *m*, respectively.

Network modules significantly dysregulated during CRC progression. The 77 network modules showing a significant dysregulation during CRC progression were taken from our previous work [61], in which we devised a computational method that combines quantitative proteomic profiling of TCGA CRC samples, protein interaction network and statistical analysis to identify significantly dysregulated cellular functions during cancer progression.

Statistical analyses. All statistical tests were performed using the R statistical environment (http://cran.r-project.org/). Unless otherwise specified, enrichments were assessed using a one-sided Fisher's exact test and P-values were corrected for multiple testing with the Benjamini–Hochberg procedure applying a significance threshold equal to 0.025. Distributions were compared using either a two-sided Mann-Whitney U test or a two-sided Kolmogorov-Smirnov test (significance threshold=0.05).

Funding

This work was partially supported by the French 'Plan Cancer 2009–2013' program (Systems Biology call, A12171AS) and the A*MIDEX project (n° ANR-11-IDEX-0001-02) funded by the «Investissements d'Avenir» French Government program, managed by the French National Research Agency (ANR), to CB. The funding organizations had no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Author contributions

AZ conceived the study, designed and performed the experiments, analyzed the data and wrote the manuscript. SB performed the experiments. LS performed the experiments and analyzed the data. CB designed the experiments, analyzed the

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the members of the TAGC laboratory for fruitful discussion and Anaïs Baudot (I2M, CNRS, France) for critically reading the manuscript.

References

1. Moore WE, Moore LV. The bacteria of periodontal diseases. Periodontol. 2000. 1994;5:66–77.

2. Han YW. Fusobacterium nucleatum: a commensal-turned pathogen. Curr. Opin. Microbiol. 2015;23C:141–7.

3. Dharmani P, Strauss J, Ambrose C, Allen-Vercoe E, Chadee K. Fusobacterium nucleatum infection of colonic cells stimulates MUC2 mucin and tumor necrosis factor alpha. Infect. Immun. 2011;79:2597–607.

4. Fardini Y, Wang X, Témoin S, Nithianantham S, Lee D, Shoham M, et al. Fusobacterium nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity. Mol. Microbiol. 2011;82:1468–80.

5. Gursoy UK, Könönen E, Uitto V-J. Intracellular replication of fusobacteria requires new actin filament formation of epithelial cells. APMIS Acta Pathol. Microbiol. Immunol. Scand. 2008;116:1063–70.

6. Han YW, Shi W, Huang GT, Kinder Haake S, Park NH, Kuramitsu H, et al. Interactions between periodontal bacteria and human oral epithelial cells: Fusobacterium nucleatum adheres to and invades epithelial cells. Infect. Immun. 2000;68:3140–6.

7. Quah SY, Bergenholtz G, Tan KS. Fusobacterium nucleatum induces cytokine production through Toll-like-receptor-independent mechanism. Int. Endod. J. 2014;47:550–9.

8. Strauss J, Kaplan GG, Beck PL, Rioux K, Panaccione R, Devinney R, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. Inflamm. Bowel Dis. 2011;17:1971–8.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22:299–306.

10. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22:292–8.

11. McCoy AN, Araújo-Pérez F, Azcárate-Peril A, Yeh JJ, Sandler RS, Keku TO. Fusobacterium is associated with colorectal adenomas. PloS One. 2013;8:e53653.

12. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15:382–92.

13. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013;14:207–15.

14. Mima K, Nishihara R, Qian ZR, Cao Y, Sukawa Y, Nowak JA, et al. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. Gut. 2015;

15. Mima K, Sukawa Y, Nishihara R, Qian ZR, Yamauchi M, Inamura K, et al. Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. JAMA Oncol. 2015;1:653–61.

16. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe. 2013;14:195–206.

17. Han YW, Ikegami A, Rajanna C, Kawsar HI, Zhou Y, Li M, et al. Identification and characterization of a novel adhesin unique to oral fusobacteria. J. Bacteriol. 2005;187:5330–40.

18. Kaplan CW, Ma X, Paranjpe A, Jewett A, Lux R, Kinder-Haake S, et al. Fusobacterium nucleatum outer membrane proteins Fap2 and RadD induce cell death in human lymphocytes. Infect. Immun. 2010;78:4773–8.

19. Gur C, Ibrahim Y, Isaacson B, Yamin R, Abed J, Gamliel M, et al. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. Immunity. 2015;42:344–55.

20. Elde NC, Malik HS. The evolutionary conundrum of pathogen mimicry. Nat. Rev. Microbiol. 2009;7:787–97.

21. Davey NE, Travé G, Gibson TJ. How viruses hijack cell regulation. Trends

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Biochem. Sci. 2011;36:159–69.

22. Via A, Uyar B, Brun C, Zanzoni A. How pathogens use linear motifs to perturb host cell networks. Trends Biochem. Sci. 2015;40:36–48.

23. Ludin P, Nilsson D, Mäser P. Genome-wide identification of molecular mimicry candidates in parasites. PloS One. 2011;6:e17546.

24. Baxt LA, Garza-Mayers AC, Goldberg MB. Bacterial subversion of host innate immune pathways. Science. 2013;340:697–701.

25. Rudel T, Kepp O, Kozjak-Pavlovic V. Interactions between bacterial pathogens and mitochondrial cell death pathways. Nat. Rev. Microbiol. 2010;8:693–705.

26. Haglund CM, Welch MD. Pathogens and polymers: microbe-host interactions illuminate the cytoskeleton. J. Cell Biol. 2011;195:7–17.

27. Uetz P, Dong Y-A, Zeretzke C, Atzler C, Baiker A, Berger B, et al. Herpesviral protein networks and their interaction with the human proteome. Science. 2006;311:239–42.

28. de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugué S, et al. Hepatitis C virus infection protein network. Mol. Syst. Biol. 2008;4:230.

29. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature. 2012;481:365–70.

30. Dyer MD, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, et al. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PloS One. 2010;5:e12089.

31. Blasche S, Arens S, Ceol A, Siszler G, Schmidt MA, Häuser R, et al. The EHEC-host interactome reveals novel targets for the translocated intimin receptor. Sci. Rep. 2014;4:7531.

32. Mirrashidi KM, Elwell CA, Verschueren E, Johnson JR, Frando A, Von Dollen J, et al. Global Mapping of the Inc-Human Interactome Reveals that Retromer Restricts Chlamydia Infection. Cell Host Microbe. 2015;18:109–21.

33. Weß ling R, Epple P, Altmann S, He Y, Yang L, Henz SR, et al. Convergent Targeting of a Common Host Protein-Network by Pathogen Effectors from Three Kingdoms of Life. Cell Host Microbe. 2014;16:364–375.

34. Ahn H-J, Kim S, Kim H-E, Nam H-W. Interactions between secreted GRA proteins and host cell proteins across the paratitophorous vacuolar membrane in the parasitism of Toxoplasma gondii. Korean J. Parasitol. 2006;44:303–12.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

35. Wu H-J, Wang AH-J, Jennings MP. Discovery of virulence factors of pathogenic bacteria. Curr. Opin. Chem. Biol. 2008;12:93–101.

36. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. Cell. 2011;144:986–98.

37. McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, et al. Computational prediction of type III and IV secreted effectors in gram-negative bacteria. Infect. Immun. 2011;79:23–32.

38. Wang Y, Wei X, Bao H, Liu S-L. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.

39. Garamszegi S, Franzosa EA, Xia Y. Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. PLoS Pathog. 2013;9:e1003778.

40. Ruhanen H, Hurley D, Ghosh A, O'Brien KT, Johnston CR, Shields DC. Potential of known and short prokaryotic protein motifs as a basis for novel peptide-based antibacterial therapeutics: a computational survey. Front. Microbiol. 2014;5:4.

41. Arnold R, Boonen K, Sun MGF, Kim PM. Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. Methods San Diego Calif. 2012;57:508–18.

42. Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, Lykidis A, et al. Genome sequence and analysis of the oral bacterium Fusobacterium nucleatum strain ATCC 25586. J. Bacteriol. 2002;184:2005–18.

43. Desvaux M, Khan A, Beatson SA, Scott-Tucker A, Henderson IR. Protein secretion systems in Fusobacterium nucleatum: genomic identification of Type 4 piliation and complete Type V pathways brings new insight into mechanisms of pathogenesis. Biochim. Biophys. Acta. 2005;1713:92–112.

44. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. 2011;8:785–6.

45. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol. 2005;5:58.

46. Kaplan A, Kaplan CW, He X, McHardy I, Shi W, Lux R. Characterization of aid1, a novel gene involved in Fusobacterium nucleatum interspecies interactions. Microb. Ecol. 2014;68:379–87.

47. Marín M, Uversky VN, Ott T. Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race. Plant Cell.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2013;25:3153–7.

48. Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, Uversky VN, et al. Structural disorder in viral proteins. Chem. Rev. 2014;114:6880–911.

49. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2012;40:D641-645.

50. Dean P. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. FEMS Microbiol. Rev. 2011;35:1100–25.

51. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol. Biosyst. 2012;8:268–81.

52. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290-301.

53. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, et al. ELM–the database of eukaryotic linear motifs. Nucleic Acids Res. 2012;40:D242–51.

54. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–114.

55. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691-697.

56. Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C. Extreme multifunctional proteins identified from a human protein interaction network. Nat. Commun. 2015;6:7412.

57. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52.

58. Becker E, Robisson B, Chapple CE, Guénoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinforma. Oxf. Engl. 2012;28:84–90.

59. Chapple CE, Brun C. Redefining protein moonlighting. Oncotarget. 2015;6:16812–3.

60. Manson McGuire A, Cochrane K, Griggs AD, Haas BJ, Abeel T, Zeng Q, et al. Evolution of invasion in a diverse set of Fusobacterium species. mBio. 2014;5:e01864.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

61. Zanzoni A, Brun C. Integration of quantitative proteomics data and interaction networks: Identification of dysregulated cellular functions during cancer progression. Methods San Diego Calif. 2016;93:103–9.

62. Tyagi N, Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between Helicobacter pylori and a human host. Mol. Biosyst. 2009;5:1630–5.

63. Doolittle JM, Gomez SM. Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS Negl. Trop. Dis. 2011;5:e954.

64. Schleker S, Garcia-Garcia J, Klein-Seetharaman J, Oliva B. Prediction and comparison of Salmonella-human and Salmonella-Arabidopsis interactomes. Chem. Biodivers. 2012;9:991–1018.

65. Kumar A, Thotakura PL, Tiwary BK, Krishna R. Target identification in Fusobacterium nucleatum by subtractive genomics approach and enrichment analysis of host-pathogen protein-protein interactions. BMC Microbiol. 2016;16:84.

66. Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Med. Genomics. 2009;2:27.

67. Hagai T, Azia A, Babu MM, Andino R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Rep. 2014;7:1729–39.

68. Christie-Oleza JA, Piña-Villalonga JM, Bosch R, Nogales B, Armengaud J. Comparative proteogenomics of twelve Roseobacter exoproteomes reveals different adaptive strategies among these marine bacteria. Mol. Cell. Proteomics MCP. 2012;11:M111.013110.

69. Tjalsma H, Lambooy L, Hermans PW, Swinkels DW. Shedding & shaving: disclosure of proteomic expressions on a bacterial face. Proteomics. 2008;8:1415–28.

70. Pérez-Cruz C, Delgado L, López-Iglesias C, Mercade E. Outer-inner membrane vesicles naturally secreted by gram-negative pathogenic bacteria. PloS One. 2015;10:e0116896.

71. Chodavarapu S, Felczak MM, Kaguni JM. Two forms of ribosomal protein L2 of Escherichia coli that inhibit DnaA in DNA replication. Nucleic Acids Res. 2011;39:4180–91.

72. Edwards RJ, Palopoli N. Computational prediction of short linear motifs from protein sequences. Methods Mol. Biol. Clifton NJ. 2015;1268:89–141.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

73. Edwards RJ, Davey NE, O'Brien K, Shields DC. Interactome-wide prediction of short, disordered protein interaction motifs in humans. Mol. Biosyst. 2012;8:282–95.

74. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.

75. Mukhtar MS, Carvunis A-R, Dreze M, Epple P, Steinbrenner J, Moore J, et al. Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science. 2011;333:596–601.

76. Navratil V, de Chassey B, Combe CR, Lotteau V. When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. BMC Syst. Biol. 2011;5:13.

77. Bierne H, Hamon M, Cossart P. Epigenetics and bacterial infections. Cold Spring Harb. Perspect. Med. 2012;2:a010272.

78. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382–7.

79. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat. Med. 2013;19:619–25.

80. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LPMH, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat. Med. 2013;19:614–8.

81. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012;40:D71-75.

82. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinforma. Oxf. Engl. 2004;20:2138–9.

83. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinforma. Oxf. Engl. 2005;21:3433–4.

84. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Struct. Lond. Engl. 1993. 2003;11:1453–9.

85. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 2015;43:W580-584.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

86. Davey NE, Haslam NJ, Shields DC, Edwards RJ. SLiMSearch 2.0: biological context for short linear motifs in proteins. Nucleic Acids Res. 2011;39:W56–60.

87. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinforma. Oxf. Engl. 2007;23:950–6.

88. Stein A, Céol A, Aloy P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2011;39:D718–23.

89. Weatheritt RJ, Jehl P, Dinkel H, Gibson TJ. iELM--a web server to explore short linear motif-mediated interactions. Nucleic Acids Res. 2012;40:W364-369.

90. Navratil V, de Chassey B, Meyniel L, Delmotte S, Gautier C, André P, et al. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. Nucleic Acids Res. 2009;37:D661-668.

91. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;

92. The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010;38:D331–5.

93. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419.

94. de Wit M, Kant H, Piersma SR, Pham TV, Mongera S, van Berkel MPA, et al. Colorectal cancer candidate biomarkers identified by tissue secretome proteome profiling. J. Proteomics. 2014;99:26–39.

95. An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database J. Biol. Databases Curation. 2014;2014:bau015.

96. Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat. Methods. 2011;8:528–9.

97. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma. Oxf. Engl. 2012;28:3150–2.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables

Table 1. Enrichment of Pfam domains in the *Fuso*Secretome compared to non-secreted proteins.

Pfam domain	Pfam Clan[a]	Function	VFDB[b]	Human[c]	# *Fuso*Secretome[d]	# non-secreted[e]	Corrected P-value[f]
MORN repeat variant	MORN repeat	Invasion	-	-	16	2	2.17×10^{-12}
Autotransporter beta-domain	Outer membrane beta-barrel protein superfamily	Adhesion	✓	-	9	0	2.4×10^{-7}
Haemolysin secretion/activation protein ShlB/FhaC/HecB	Outer membrane beta-barrel protein superfamily	Secretion	✓	-	5	0	0.002
TonB-dependent Receptor Plug Domain	Ubiquitin superfamily	Transport	✓	-	5	0	0.002
TonB dependent receptor	Outer membrane beta-barrel protein superfamily	Transport	✓	-	4	0	0.011
Surface antigen variable number repeat	POTRA domain superfamily	Folding	-	-	4	0	0.011
YadA-like C-terminal region	Pilus subunit	Adhesion	✓	-	4	0	0.011
Haemagluttinin repeat	Pectate lyase-like beta helix	Adhesion	✓	-	4	0	0.011
Bacterial extracellular solute-binding proteins, family 5 Middle	Periplasmic binding protein clan	Transport	-	-	5	4	0.082
Coiled stalk of trimeric autotransporter adhesion	-	Adhesion	✓	-	3	0	0.09
Pyruvate flavodoxin/ferredoxin oxidoreductase, thiamine diP-bdg	Thiamin diphosphate-binding superfamily	Metabolism	-	-	3	0	0.09
TPR repeat	Tetratrico peptide repeat superfamily	Protein binding	✓	✓	6	8	0.094

[a] A clan is defined as a collection of related Pfam entries sharing sequence or structural similarity. [b] Pfam entry detected in at least one protein sequence stored in the database of known bacterial virulence factors. [c] Pfam entry detected in at least one human protein sequence. [d] Number of occurrences in the *Fuso*Secretome. [e] Number of occurrences in non-secreted proteins.
[f] Pfam domain matches with a corrected P-value <0.1.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Significant Gene Ontology and pathways annotations among *Fuso*Secretome inferred human interactors.

Annotation source	Annotation ID	Annotation name	Adjusted P-value
Biological Process			
	GO:0006415	translational termination	$7.57x10^{-41}$
	GO:0006414	translational elongation	$2.65x10^{-34}$
	GO:0006457	protein folding	$2.51x10^{-29}$
	GO:0006413	translational initiation	$6.63x10^{-27}$
	GO:0072376	protein activation cascade	$1.17x10^{-25}$
	GO:0051604	protein maturation	$5.45x10^{-23}$
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$1.93x10^{-22}$
	GO:0006956	complement activation	$3.62x10^{-21}$
	GO:0002697	regulation of immune effector process	$4.03x10^{-18}$
	GO:0030449	regulation of complement activation	$4.10x10^{-18}$
Cellular Component			
	GO:0005840	ribosome	$9.81x10^{-37}$
	GO:0022626	cytosolic ribosome	$2.76x10^{-23}$
	GO:0005912	adherens junction	$4.15x10^{-22}$
	GO:0098552	side of membrane	$2.46x10^{-18}$
	GO:0030055	cell-substrate junction	$2.85x10^{-17}$
	GO:0072562	blood microparticle	$9.20x10^{-17}$
	GO:0005761	mitochondrial ribosome	$4.28x10^{-12}$
	GO:0019897	extrinsic component of plasma membrane	$2.35x10^{-09}$
	GO:0005911	cell-cell junction	$4.54x10^{-08}$
	GO:0031012	extracellular matrix	$3.91x10^{-07}$
KEGG			
	KEGG:03010	Ribosome	$4.34x10^{-39}$
	KEGG:04610	Complement and coagulation cascades	$2.28x10^{-22}$
	KEGG:04514	Cell adhesion molecules (CAMs)	$2.88x10^{-09}$
	KEGG:04141	Protein processing in endoplasmic reticulum	$2.03x10^{-08}$
	KEGG:04660	T cell receptor signaling pathway	$1.33x10^{-07}$
	KEGG:05150	Staphylococcus aureus infection	$1.78x10^{-06}$
	KEGG:04380	Osteoclast differentiation	$7.25x10^{-06}$
	KEGG:05203	Viral carcinogenesis	$8.21x10^{-06}$
	KEGG:05169	Epstein-Barr virus infection	$1.11x10^{-05}$
	KEGG:05164	Influenza A	$1.77x10^{-05}$
Reactome			
	REAC:1592389	Activation of Matrix Metalloproteinases	$4.77x10^{-28}$
	REAC:192823	Viral mRNA Translation	$2.14x10^{-19}$
	REAC:156902	Peptide chain elongation	$2.14x10^{-19}$
	REAC:975956	Nonsense Mediated Decay independent of the Exon Junction Complex	$1.83x10^{-17}$
	REAC:977606	Regulation of Complement cascade	$5.13x10^{-16}$
	REAC:202733	Cell surface interactions at the vascular wall	$7.22x10^{-16}$
	REAC:5368287	Mitochondrial translation	$3.61x10^{-14}$
	REAC:3371453	Regulation of HSF1-mediated heat shock response	$3.92x10^{-14}$
	REAC:3371599	Defective HLCS causes multiple carboxylase deficiency	$1.33x10^{-08}$
	REAC:420597	Nectin/Necl trans heterodimerization	$1.33x10^{-08}$

For each annotation source, the ten most significant terms are reported. The full list of annotation enrichments is available in Table S6.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3. Network module significantly enriched in inferred human interactors.

Module	Module proteins	Interactors	Inferred interactions	*Fuso*Secretome proteins	Annotations
9	74	13	188	64	immune response-regulating cell surface receptor signaling pathway (GO:0002768), cell-cell junction (GO:0005911)
16	67	12	122	65	metal ion homeostasis (GO:0055065), cell surface (GO:0009986)
19	80	15	260	55	cellular response to organonitrogen compound (GO:0071417), membrane raft (GO:0045121)
42	140	19	312	56	endocytosis (GO:0006897), membrane raft (GO:0045121)
74	113	26	78	26	extracellular structure organization (GO:0043062), cell surface (GO:0009986)
78	292	54	752	100	immune response-regulating cell surface receptor signaling pathway (GO:0002768), cell surface (GO:0009986)
89	45	10	169	48	immune response-activating cell surface receptor signaling pathway (GO:0002429), nucleolar ribonuclease P complex (GO:0005655)
90	51	10	28	12	I-kappaB kinase/NF-kappaB cascade (GO:0007249), inclusion body (GO:0016234)
138	126	22	57	26	I-kappaB kinase/NF-kappaB cascade (GO:0007249), perinuclear region of cytoplasm (GO:0048471)
165	81	17	327	71	neuron projection guidance (GO:0097485), synapse (GO:0045202)
194	38	12	119	48	G1/S transition of mitotic cell cycle (GO:0000082), cyclin-dependent protein kinase holoenzyme complex (GO:0000307)
216	47	11	261	80	blood coagulation (GO:0007596), membrane raft (GO:0045121)
246	50	15	83	32	T cell activation (GO:0042110), Golgi membrane (GO:0000139)
277	25	7	7	2	collagen catabolic process (GO:0030574), extracellular matrix (GO:0031012)
298	106	16	242	75	actin cytoskeleton organization (GO:0030036), Arp2/3 protein complex (GO:0005885)
300	37	9	164	49	stress-activated MAPK cascade (GO:0051403), nuclear speck (GO:0016607)
371	25	7	79	48	actin filament organization (GO:0007015), lamellipodium (GO:0030027)
433	38	9	121	49	positive regulation of intracellular protein kinase cascade (GO:0010740), spindle (GO:0005819)
451	40	9	55	34	mitotic cell cycle phase transition (GO:0044772), heterochromatin (GO:0000792)
456	36	9	142	59	regulation of system process (GO:0044057), dendrite (GO:0030425)
563	26	8	43	29	regulation of sequence-specific DNA binding transcription factor activity (GO:0051090), external side of plasma membrane (GO:0009897)
571	42	11	206	53	cell cycle phase transition (GO:0044770), transcription factor complex (GO:0005667)
577	33	10	66	48	complement activation (GO:0006956), ER membrane insertion complex (GO:0072379)
587	17	6	109	46	axonogenesis (GO:0007409), signalosome (GO:0008180)
615	36	11	26	10	response to unfolded protein (GO:0006986), perinuclear region of cytoplasm (GO:0048471)
625	25	8	226	104	regulation of sequence-specific DNA binding transcription factor activity (GO:0051090), chromatin (GO:0000785)
689	38	14	151	72	blood coagulation (GO:0007596), apical junction complex (GO:0043296)
702	23	8	157	46	peptidyl-tyrosine phosphorylation (GO:0018108), nucleolar ribonuclease P complex (GO:0005655)
745	18	6	113	46	axon guidance (GO:0007411), cell leading edge (GO:0031252)
794	22	7	151	46	gamma-aminobutyric acid signaling pathway (GO:0007214), postsynaptic membrane (GO:0045211)
831	15	6	129	45	Fc-gamma receptor signaling pathway involved in phagocytosis (GO:0038096), cell leading edge (GO:0031252

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For each module the following information is reported: identifier, number of constituent proteins, number of inferred human interactors in the module, number of inferred interactions between proteins in the module and *Fuso*Secretome proteins, number of interacting *Fuso*Secretome proteins, representative annotations (Biological Process and Cellular Component) selected as the most frequent and significantly enriched annotations for the given module (for the complete list of functional annotations see Table S9).

Table 4. List of candidate *Fuso*Secretome virulence factors.

UniprotKB AC	Protein name	Gene symbol	Domains	Interacting Domains	Interacting SLiMs
Q8RIM1	Fusobacterium outer membrane protein family	FN1554	Autotransporter[a]	-	LIG_FHA_1, LIG_FHA_2, LIG_PP1[b], LIG_SH2_SRC[b], LIG_SH2_STAT5, LIG_SH3_3[b] LIG_SUMO_SBM_1[b], MOD_N-GLC_1[b], TRG_ENDOCYTIC_2[b]
Q8RGK2	Hemolysin	FN0291	Fil_haemagg_2[a]	-	LIG_FHA_1, LIG_FHA_2, LIG_Rb_pABgroove_1[b], LIG_SH2_GRB2, LIG_SH2_SRC[b], LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_N-GLC_1[b] MOD_PIKK_1, TRG_ENDOCYTIC_2[b]
Q8RGT9	Peptide methionine sulfoxide reductase MsrA	msrA	PMSR, SelR	PMSR, SelR	CLV_PCSK_PC1ET2_1, LIG_FHA_1, LIG_SH2_GRB2, LIG_SH2_SRC[b], LIG_SH2_STAT5, MOD_Cter_Amidation, MOD_PIKK_1, TRG_ENDOCYTIC_2[b]
Q8RHB9	Hypothetical exported 24-amino acid repeat protein	FN2118	MORN_2	-	LIG_SH2_GRB2, LIG_SH2_SRC[b]
Q8R609	Pyruvate-flavodoxin oxidoreductase	FN1421	POR_N, POR, EKR, Fer4_7, TPP_enzyme_C	TPP_enzyme_C	LIG_CYCLIN_1[b], LIG_SH2_GRB2, LIG_SH2_STAT5, LIG_SH3_3[b], LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_PIKK_1, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]
Q8RH03	Chaperone protein DnaJ	dnaJ	DnaJ, DnaJ_CXXCXGXG, CTDII	DnaJ, DnaJ_CXXCXGXG	CLV_NDR_NDR_1, CLV_PCSK_SKI1_1, LIG_CYCLIN_1[b], LIG_FHA_2, LIG_SH2_STAT5, LIG_SH3_3[b], LIG_SUMO_SBM_1[b], LIG_TRAF2_1[b], LIG_WW_Pin1_4, MOD_CK2_1[b] MOD_PLK, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]
Q8R643	Pyruvate-flavodoxin oxidoreductase	FN1170	POR_N, POR, EKR, Fer4_7, TPP_enzyme_C	TPP_enzyme_C	LIG_BRCT_BRCA1_1, LIG_SH2_GRB2, LIG_SH2_STAT5, LIG_SH3_3[b], LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]
Q8RDQ9	Fusobacterium outer membrane protein family	FN1449	-	-	CLV_PCSK_SKI1_1, LIG_FHA_1, LIG_FHA_2, LIG_PDZ_Class_2, LIG_SH2_SRC[b], LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], MOD_GSK3_1, MOD_N-GLC_1[b], MOD_PLK
Q8R608	Serine protease	FN1426	Peptidase_S8[a], Autotrns_rpt[a], Autotransporte[a]r	Peptidase_S8	CLV_PCSK_SKI1_1, LIG_FHA_2, LIG_PDZ_Class_2, LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_PKA_2, TRG_ENDOCYTIC_2[b]
Q8RFV3	Hypothetical cytosolic protein	FN0579	MG1, A2M_N, A2M_N_2, A2M	A2M_N, A2M_N_2, A2M	LIG_CYCLIN_1[b], LIG_FHA_2, LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], MOD_CK2_1[b], MOD_PIKK_1, MOD_PKA_2
Q8R5P1	DNAse I homologous protein DHP2	FN0891	Exo_endo_phos	Exo_endo_phos	LIG_SH2_GRB2, TRG_ENDOCYTIC_2[b]
Q8R5Y8	Biotin carboxyl carrier protein of glutaconyl-COA decarboxylase	FN0200	Biotin_lipoyl	Biotin_lipoyl	LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_ProDKin_1, MOD_SUMO[b]
Q8R6D6	Serine protease	FN1950	Peptidase_S8[a], Autotransporter[a]	Peptidase_S8	LIG_FHA_2, LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_PKA_2

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Q8RE26	Single-stranded DNA-binding protein	ssb	SSB[a]	SSB	LIG_BRCT_BRCA1_1, LIG_FHA_1, LIG_FHA_2 LIG_PDZ_Class_2, LIG_SUMO_SBM_1[b], MOD_PKA_2, TRG_ENDOCYTIC_2[b]
Q8REJ1	Dipeptide-binding protein	FN1111	SBP_bac_5	-	CLV_PCSK_SKI1_1, LIG_BRCT_BRCA1_1, LIG_CYCLIN_1[b], LIG_SH2_GRB2, LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK2_1[b], MOD_ProDKin_1
Q8RG20	Hemin receptor	FN0499	Plug[a]	-	CLV_PCSK_PC1ET2_1, LIG_CYCLIN_1[b], LIG_FHA_1, LIG_MAPK_1, LIG_PDZ_Class_2, LIG_SH2_STAT5, LIG_SH3_3[b], LIG_SUMO_SBM_1[b], MOD_GSK3_1, MOD_N-GLC_1[b]
Q8RGC9	Iron(III)-binding protein	FN0375	SBP_bac_8	-	LIG_PP1[b], LIG_SH2_GRB2, LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK1_1, MOD_ProDKin_1
Q8RGZ4	Hemolysin activator protein	FN0131	POTRA_2[a], ShlB[a]	-	CLV_PCSK_PC1ET2_1, LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], LIG_TRAF2_1[b]
Q8RH05	Chaperone protein DnaK	dnaK	HSP70	HSP70	CLV_PCSK_PC1ET2_1, CLV_PCSK_SKI1_1, LIG_BRCT_BRCA1_1, LIG_EVH1_1, LIG_FHA_2, LIG_SH2_STAT5, LIG_SH3_3[b], LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_PIKK_1, MOD_PLK, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b], TRG_LysEnd_APsAcLL_1[b]
Q8RHM4	Tetratricopeptide repeat protein	FN1990	DnaJ	DnaJ	-
Q8RHU4	Hypothetical lipoprotein	FN1899	DUF3798	-	CLV_C14_Caspase3-7, LIG_FHA_1, LIG_FHA_2 LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK2_1[b], MOD_GSK3_1 MOD_PIKK_1, MOD_PLK, MOD_ProDKin_1
Q8RI19	Hemolysin	FN1817	Fil_haemagg_2[a]	-	LIG_CYCLIN_1[b], LIG_FHA_1, LIG_FHA_2, LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK1_1, MOD_CK2_1[b], MOD_GSK3_1, MOD_N-GLC_1[b], MOD_PIKK_1 MOD_PLK, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]
Q8RI47	Tetratricopeptide repeat family protein	FN1787	TPR_11[a], TPR_1	TPR_11, TPR_1	-
Q8RIF8	50S ribosomal protein L2	rplB	Ribosomal_L2, Ribosomal_L2_C	Ribosomal_L2, Ribosomal_L2_C	CLV_PCSK_SKI1_1, LIG_BIR_II_1, LIG_PP1[b], LIG_SH3_3[b], LIG_SUMO_SBM_1[b], LIG_USP7_1[b], LIG_WW_Pin1_4, MOD_PLK, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]
Q8RII5	Competence protein	FN1611	HHH_3	-	LIG_TRAF2_1[b], MOD_CK2_1[b]
Q8RIK0	Hypothetical lipoprotein	FN1590	DUF3798	-	CLV_C14_Caspase3-7, LIG_FHA_1, LIG_FHA_2 LIG_SH2_STAT5, LIG_SUMO_SBM_1[b], LIG_WW_Pin1_4, MOD_CK2_1[b], MOD_GSK3_1 MOD_PIKK_1, MOD_PLK, MOD_ProDKin_1, TRG_ENDOCYTIC_2[b]

For every protein, the detected Pfam domains are reported along with the list of domains and SLiMs for which at least one interaction has been inferred. [a] Pfam entry detected in at least one protein sequence stored in the database of known bacterial virulence factors. [b] Motif for which it was experimentally identified at least one instance in a viral or bacterial protein.

bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

Figure 1. Flow strategy of our computational approach**.**

Figure 2. Disorder propensity of the *Fuso*Secretome**.** Putative secreted proteins show a significantly higher fraction of disordered residues compared to non-secreted proteins (P-value = 1.9×10^{-4}, Kolgomorov-Smirnov test, two-sided). It is to note that SecretomeP exploits disorder as a predicting feature. Therefore, to avoid circularity issues, we assessed the disorder propensity of the 61 proteins identified by the SignalP algorithm only.

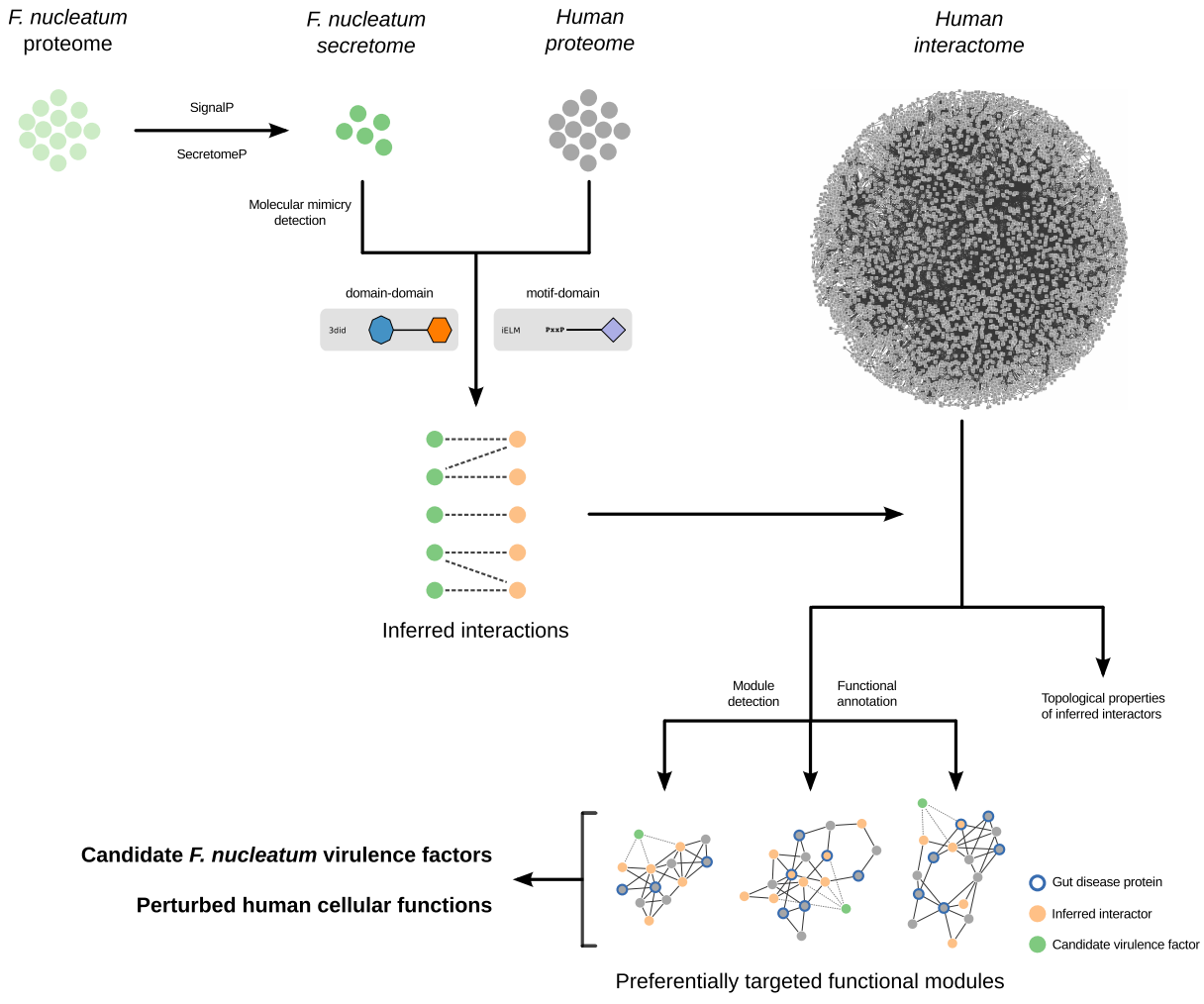
Figure 3. Topological properties of inferred human interactors in the human interactome. (A) Inferred human interactors have more interaction partners and (B) higher values of betwenness centrality compared to non-interacting proteins in the human interactome.

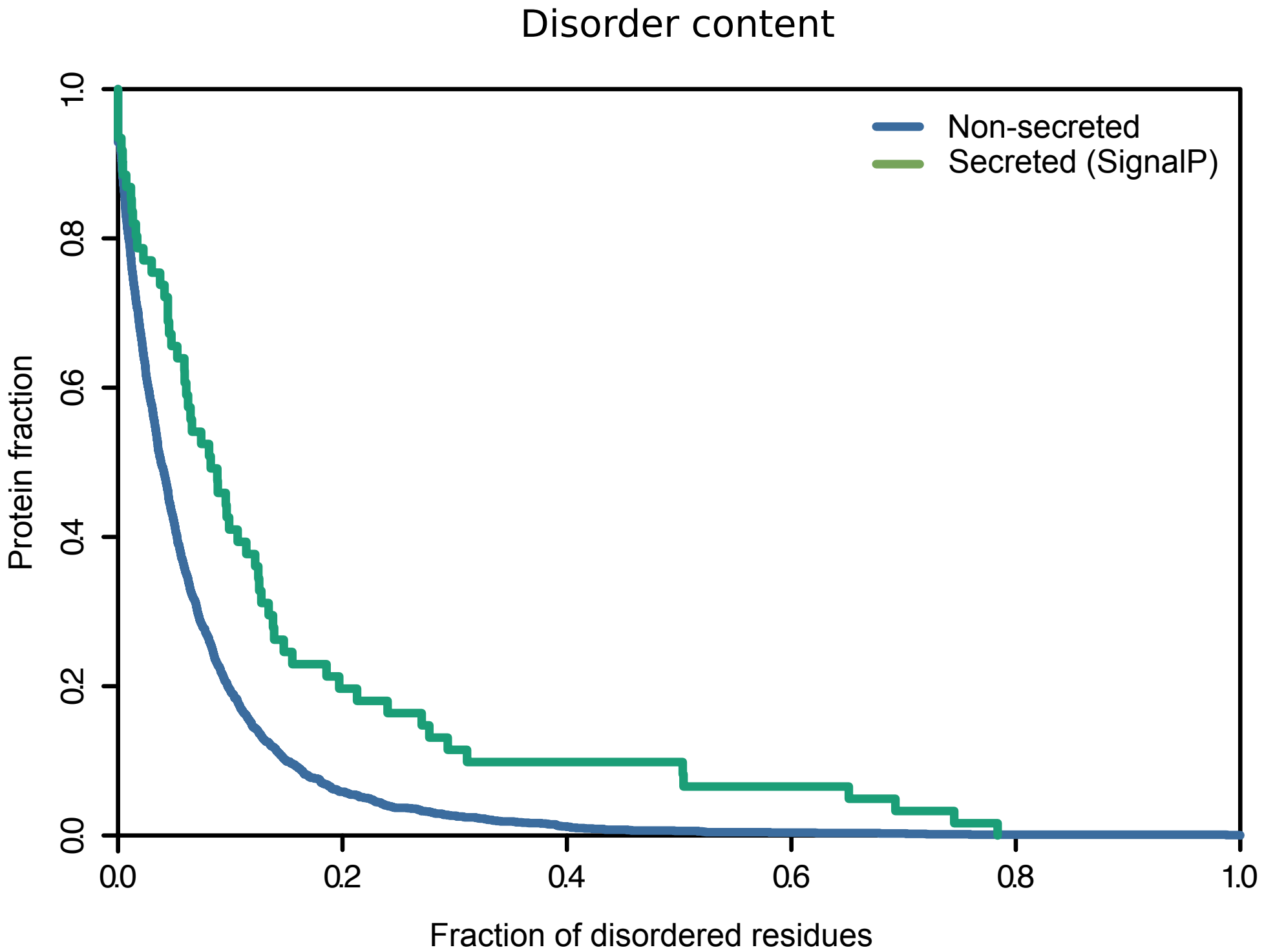
Figure 4. Interaction network between *Fuso*Secretome candidate virulence factors and preferentially targeted modules. (A) Virulence factors are depicted as green rectangular nodes labelled with respective gene symbol, whereas network modules as orange circles, whose size is proportional to the number of proteins belonging to each module and the label number corresponds the module identifier. Edge width is proportional to the number of inferred interactions of a virulence factor with a given module. Network modules enriched in gut disease proteins are highlighted with a violet border (see main text for details). (B) The protein Fap2 (FN1449) interacts with 9 proteins (orange nodes with a green border) of Module 9 and (C) the MORN2 domain containing protein (FN2118) interacts with 8 proteins in Module 89. Both modules contain proteins involved in immune response.

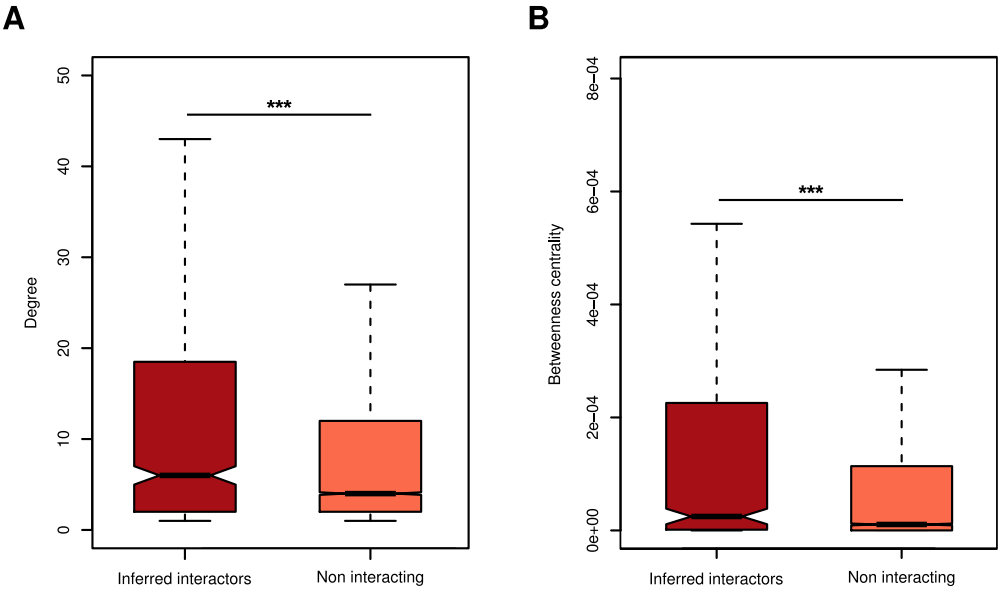
bioRxiv preprint doi: https://doi.org/10.1101/094136; this version posted December 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

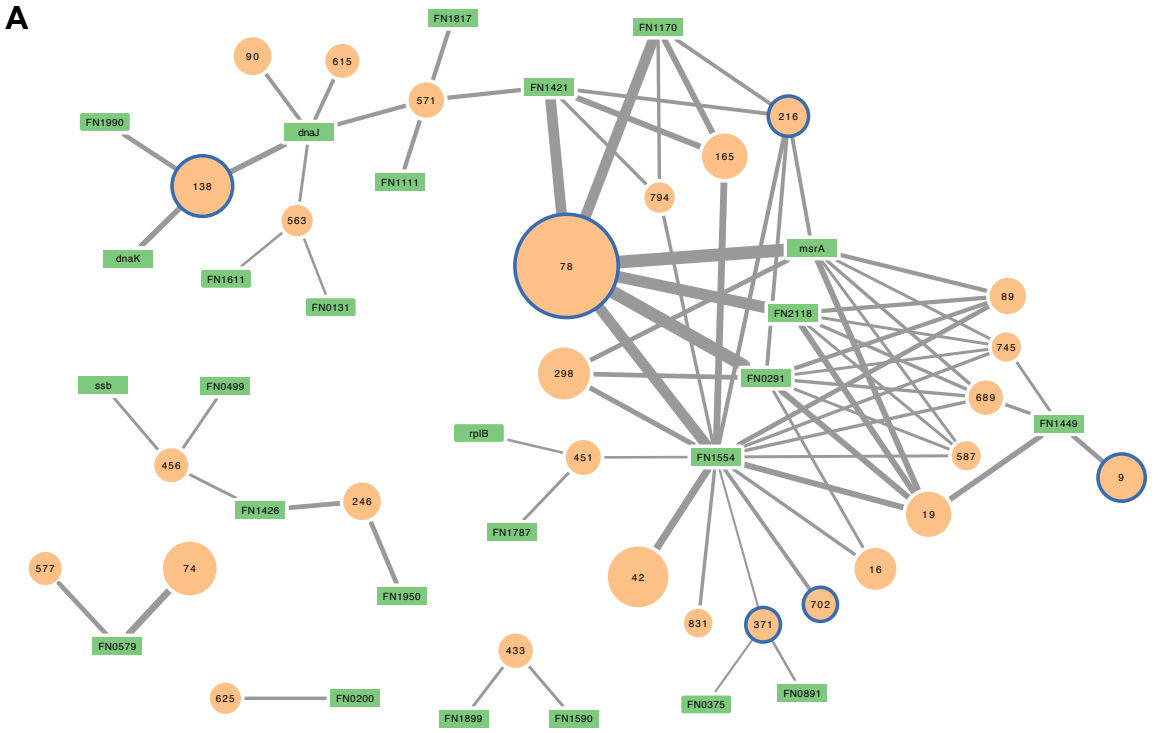
Figure 5. Enrichment of *Fuso*Secretome inferred human interactors and gut disease related proteins in network modules. Each column of the heatmap represents a module. The color of the cells corresponds to the log-transformed enrichment ratio. Pink circles indicate enriched sets. Modules showing a significant dysregulation in CRC progression are highlighted by an empty circle with green border. For the six modules showing an enrichment in inferred interactions and at least in one of gut disease related proteins, the most representative functions are reported. FusoExpr: genes whose expression correlates with *F. nucleatum* abundance in CRC patients; CRC: genes mutated in colorectal cancer samples; IBD: genes associated to inflammatory bowel's disease; CD and UC: genes specifically associated to Crohn's disease and ulcerative colitis respectively.

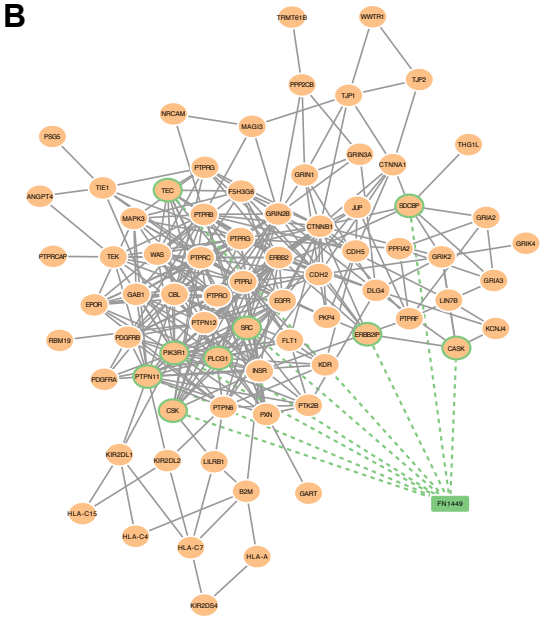
Supplementary Figure 1. Assessment of the disorder propensity of the *F.nucleatum* secreted proteins with additional prediction algorithms.

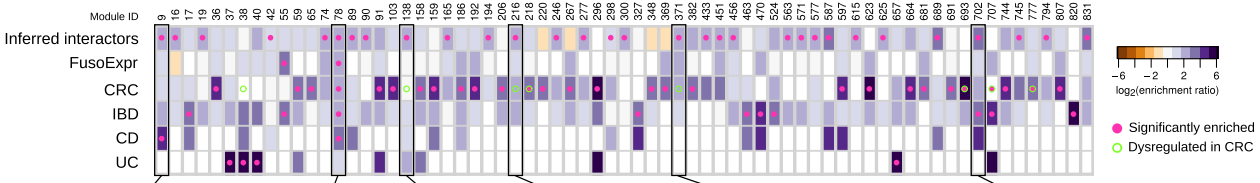

F. nucleatum proteome
F. nucleatum secretome
Human proteome
Human interactome
SignalP
SecretomeP
Molecular mimicry detection
domain-domain
motif-domain
3did
iELM
PxxP
Inferred interactions
Module detection
Functional annotation
Topological properties of inferred interactors
Candidate F. nucleatum virulence factors
Perturbed human cellular functions
Gut disease protein
Inferred interactor
Candidate virulence factor
Preferentially targeted functional modules


Disorder content
Non-secreted
Secreted (SignalP)
Protein fraction
Fraction of disordered residues
0.0
0.2
0.4
0.6
0.8
1.0


A
Degree
0
10
20
30
40
50
***
Inferred interactors
Non interacting
B
Betweenness centrality
0e+00
2e−04
4e−04
6e−04
8e−04
***
Inferred interactors
Non interacting


A
FN1817
FN1170
90
615
571
FN1421
216
FN1990
dnaJ
165
138
FN1111
794
563
dnaK
78
msrA
FN1611
FN0131
89
FN2118
745
298
FN0291
689
ssb
FN0499
FN1449
rplB
451
FN1554
587
9
456
FN1426
246
19
FN1787
577
74
42
16
FN1950
702
831
371
FN0579
433
625
FN0200
FN0891
FN0375
FN1899
FN1590


B
TRMT61B
WWTR1
PPP2CB
TJP1
TJP2
NRCAM
MAGI3
PSG5
GRINA
THG1L
CTNNA1
PTPRG
GRIN1
ANGPT4
TIE1
TEC
FSH3G6
SDC2P
JUP
GRIN2B
GRIA2
MAPK3
PTPRB
CTNNB1
PTPRG
GRIK4
PTPRCAP
TEK
WAS
PTPRC
ERBB2
CDH5
PPFIA2
GRIK2
CDH2
GRIA3
PTPRJ
DLG4
EPOR
GAB1
CBL
PTPRD
EGFR
LIN7B
PTPN12
SFC
PKP4
PTPRF
RBM19
PDGFRB
ERBB2P
KCNJ4
FLT1
PIK3R1
PLCG1
CASK
INSR
KDR
PDGFRA
PTPN11
CSK
PTPN6
PTK2B
PXN
KIR2DL1
KIR2DL2
LILRB1
B2M
GART
FN1449
HLA-C15
HLA-C4
HLA-C7
HLA-A
KIR2DS4
C
TMEM101
PTPRN2
SMARCD3
SMARCC1
CYP4F8
ALS2CR12
B7WP10
DBNL
LAT
MAP4K1
CKAP5
FANCA
DAP3
SYK
GAB2
PLCG1
PIK3R1
SHANK3
TOM1L1
ZAP70
GRB2
SFC
NCOA1
SLK
MTTP
CCDC47
CBL
FYN
NCK1
LCK
ABL1
CRK
DLGAP2
CELSR2
FASLG
SHROOM3
CCKBR
WIPF1
GSTZ1
FVB
FN2118
RPP38
RPP40
FASTK
RPP30
POP7



Module 9

immune response-regulating cell surface receptor signaling pathway (GO:0002768)

cell-cell junction (GO:0005911)

Adherens junction (KEGG:04520)

Signaling in immune system (REAC:168256)

Module 78

immune response-regulating cell surface receptor signaling pathway (GO:0002768)

cell surface (GO:0009986)

PI3K-Akt signaling pathway (KEGG:04151)

Hemostasis (REAC:109582)

Module 138

I-kappaB kinase/NF-kappaB cascade (GO:0007249)

perinuclear region of cytoplasm (GO:0048471)

Epstein-Barr virus infection (KEGG:05169)

Toll Like Receptor 3 (TLR3) Cascade (REAC:168164)

Module 216

blood coagulation (GO:0007596)

membrane raft (GO:0045121)

Pathways in cancer (KEGG:05200)

Integrin cell surface interactions (REAC:216083)

Module 371

actin filament organization (GO:0007015)

lamellipodium (GO:0030027)

Regulation of actin cytoskeleton (KEGG:04810)

Recruitment of Grb2 to pFADK1 (REAC:354087)

Module 702

peptidyl-tyrosine phosphorylation (GO:0018108)

nucleolar ribonuclease P complex (GO:0005655)

Jak-STAT signaling pathway(KEGG:04630)

Cell surface interactions at the vascular wall (REAC:202733)