

**On the conservation and distribution of
SBDS across species**

Michael Geuenich

Dr. Marina Turlakis

2018 Summer Fellows Report

Acknowledgements

I would like to thank Dr. Marina Turlakis for the opportunity to work on this project, her ongoing support, hard work, mentorship and feedback.

I would also like to thank Dr. Johanna Rommens who was generous enough to share her knowledge and experience with us despite numerous technological issues.

Additional thanks go to Quest University Canada and the Research, Scholarship and Creative Works Committee for setting up the Summer Fellows program and giving me the opportunity to work on this project.

I am grateful for the assistance of Dr. Marjorie Wonham, Dr. Thor Veen and Dr. Robert Williamson for help with various issues and questions that arose in the process.

Finally, I would also like to thank my fellow lab members, Marcos Da Silva and Claire MacMurray for brainstorming sessions, their excitement and friendship.

Table of contents

1. Introduction	
1.1. Translation & the ribosome	4
1.2. Ribosome biogenesis & subunit joining	6
1.3. Significance of translation	7
2. Protein folding & structure	7
2.1. Primary structure	8
2.2. Secondary structure	9
2.3. Tertiary structure	11
2.4. Quaternary structure	12
2.5. The process of folding	12
2.6. Protein domains & complexes	12
3. Studying proteins	13
3.1. Multiple sequence alignments	14
4. Shwachman-Diamond syndrome	15
4.1. Disease causing <i>SBDS</i> mutations	16
4.2. The three dimensional structure of <i>SBDS</i>	22
5. References	27

On the conservation and distribution of *SBDS* across species

Introduction

One of the most fundamental ideas in biology is the central dogma of molecular biology. It describes a two-stage cellular process through which information flows from DNA to RNA and protein, thereby making the biomolecules necessary for life. The DNA of a cell can be thought of as a cookbook based upon which dishes, or biomolecules in the case of a cell, are made. DNA is double stranded and modular, and its most basic unit is a biochemical component known as a nucleotide. There are four different nucleotides, adenine, thymine, guanine and cytosine (often abbreviated as A, T, G and C respectively). The exact sequence of nucleotides determines the instructions of the recipe, i.e., the biomolecule. DNA is transcribed into RNA, a single stranded molecule also made up of nucleotides, through a process known as transcription. In the cookbook analogy this would be equivalent to a chef transcribing a single recipe from the cookbook onto a separate sheet of paper which will subsequently be used in the kitchen as the set of instructions to create the dish. One common type of RNA is messenger RNA (mRNA), which is used as the set of instructions for creating proteins in a process known as translation. Translation corresponds to the second step described by the central dogma and is performed by a piece of molecular machinery known as the ribosome. The reason the latter process is called translation is because information is moved from one language in the mRNA, that of nucleotides, to another in proteins, that of amino acids. Within the cookbook analogy, this would be equivalent to the final creation of a dish based on the copied recipe, requiring a switch from written instructions to ingredients. Proteins themselves are made up of a sequence of different kinds of biochemical components, known as amino acids, and they perform a variety of functional roles within the cell. The central dogma of molecular biology is a finely regulated process common to all of life as we know it, making it a cornerstone of basic biology. Given its importance for the functioning of all life, perturbations throughout the different stages of this process often lead to disease. Researching diseases caused by deregulation of transcription or translation thus results in a better understanding of the disease and its causes, as well as the basic biological mechanisms that are being perturbed. While a lot of research focusing on the central dogma as well as these diseases has been conducted and has resulted in a very detailed understanding of certain aspects of this process, other aspects remain less well understood.

Translation & the ribosome

Translation is performed by a piece of molecular machinery known as the ribosome. The ribosome is made up of a protein-RNA complex (RNA that makes up the ribosome is known as ribosomal RNA or rRNA), and its function is to use the instructions provided by the mRNA sequence to link together specific amino acids that will go on to form a peptide chain (the chemical bonds connecting amino acids are peptide bonds, hence the name 'peptide chain'). This peptide chain will subsequently go through downstream modifications and ultimately form a functioning protein.

The ribosome is made up of a large and a small subunit (fig. 1), between which the mRNA sequence is situated. These two subunits are held together by so called intersubunit bridges through RNA-RNA, RNA-protein and protein-protein interactions (Liiv & O'Connor, 2006). The bacterium *Thermus thermophilus* for example, has a total of 12 intersubunit bridges (Yusupov et al., 2001).

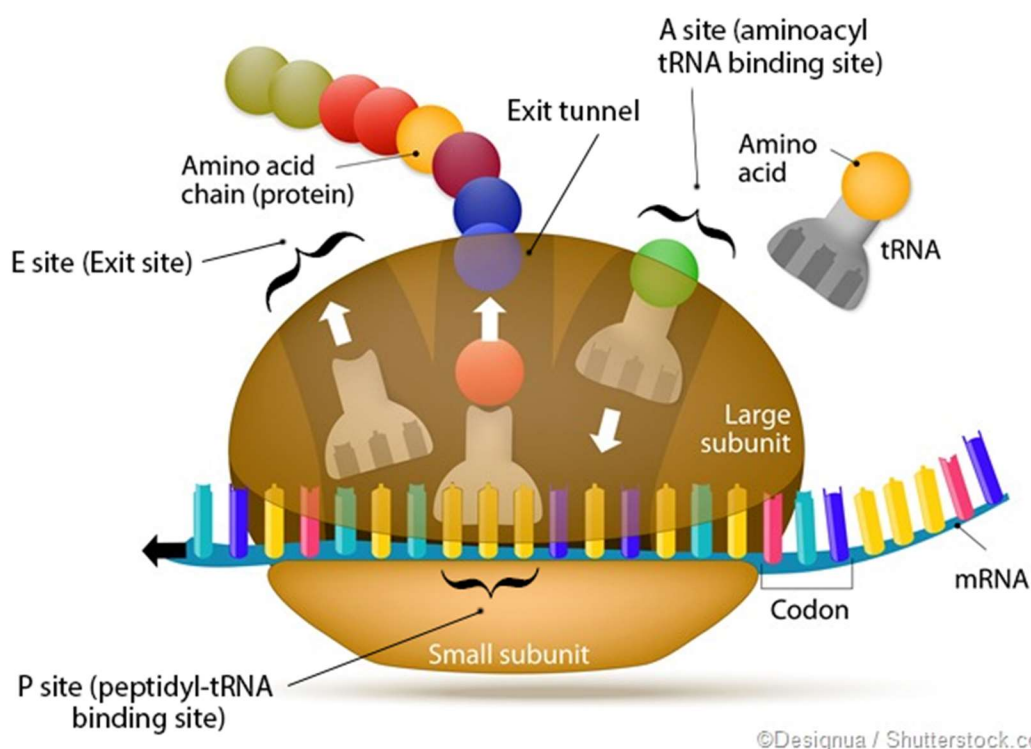


Figure 1. A depiction of a translating ribosome. Ribosomes are made up of a small and large subunit. They have three different sites, the A site, P site and E site. The mRNA sequence is fed through the ribosome and is decoded by tRNAs, which transport the different amino acids to the A site by binding to the mRNA. The ribosome then moves along the mRNA strand and connects the amino acids with peptide bonds. The polypeptide chain being created is pushed out through the exit tunnel. tRNAs detached from amino acids then move to the E site where they exit the ribosome. Image adapted from reference 1.

The mRNA sequence is read by the ribosome using a triplet code where each group of three nucleotides is known as a codon. Each codon codes for one of 20 different amino acids, or is used as a signal for the ribosome to stop translating (a so called stop codon). The amino acids are brought to the ribosome by structures composed of RNA, known as transfer RNAs (tRNAs), or as aminoacyl tRNAs (aa-tRNAs) if the amino acid is chemically bound to the tRNA. An aa-tRNA matches up three of its nucleotides (the anticodon) with a codon on the mRNA through nucleotide base pairing, a hydrogen bond mediated interaction. In this way the correct amino acid is transported to the site of peptide chain synthesis. The ribosome then catalyzes the creation of peptide bonds between the amino acids creating a polypeptide chain. Once an aa-

tRNA enters the ribosome at the A site (aminoacyl tRNA binding site) and base-pairs with the mRNA codon, the ribosome will connect the amino acid at the A site with the amino acid at the P site with a peptide bond. This occurs at the peptidyl-transferase center (PTC) contained within the large ribosomal subunit, the most conserved area of the ribosome. The PTC is an RNA enzyme that catalyzes the two main chemical reactions occurring in the ribosome, namely peptide bond formation and peptide release (Polacek & Mankin, 2005). Next, the ribosome will move along the mRNA sequence, displacing the tRNA at the A site to the P site, and the tRNA at the P site to the E site, where it will exit the ribosome (fig. 1). The growing polypeptide chain is churned out through the exit tunnel above the ribosome's P site. This process continues until the ribosome encounters a stop codon, at which point translation stops.

Ribosome biogenesis & subunit joining

To have functional translation and protein synthesis, the cell needs to correctly regulate the synthesis of ribosomes and the full suite of translation machinery (e.g. translation factors, tRNAs, etc.). The final steps of ribosome biogenesis and the start of translation require the coming together of fully functional ribosome subunits. A fully matured prokaryotic (bacteria and archaea) ribosome is denoted as a 70S ribosome, while fully matured eukaryotic (all other organisms) ribosome is denoted as an 80S ribosome. The S stands for Svedberg units and is a non-metric unit for sedimentation rate, it measures the speed at which a particle settles, the larger the value, the bigger the particle.

In eukaryotes, both ribosomal subunits are created in a part of the cell's nucleus known as the nucleolus. The large subunit is known as the 60S in eukaryotes, or 50S in prokaryotes, and the small subunit is known as the 40S in eukaryotes, or 30S in prokaryotes. After having been created, they are immature structures that need to go through a series of modifications in order to mature. Thus, they are referred to as the pre-60S and the pre-40S subunits. Both subunits will eventually be exported from the cell's nucleus into the cytoplasm where they will coalesce on a mRNA molecule in the final steps of translation initiation.

Most of our understanding of translation comes from work with simple systems, namely unicellular bacteria (e.g., *E. coli*) and yeast (e.g., *S. cerevisiae*) (Kapp & Lorsch, 2004). Through *in vivo* and *in vitro* studies in these organisms, as well as comparisons with multicellular organisms, we have learned about the increasing levels of complexity involved in the regulation of translational systems as we move from simpler forms of life to more complex multicellular forms of life such as vertebrates. One common way of studying biological processes in general is to change components of a system in a model organism. This could be done in a number of ways, some of which include deleting or modifying genes, subjecting organisms to different environmental conditions, or by dissecting an organism to examine the characteristics of different tissues. Knowledge of vertebrate and mammalian translational systems has in part been derived from studying diseases caused by disruptions in translation or ribosome biogenesis, so called ribosomopathies.

Significance of translation

Beyond being a basic biological process common to all life and thus worthwhile studying, knowledge of the ribosome and the translational process is also crucial in a clinical setting. For example, several pharmaceutical and biotechnology applications involve synthetic protein synthesis. The current method of insulin production, a peptide hormone taken by Type-I diabetics with a market size of US\$ 32 billion worldwide (Human insulin market, n.d.), relies on knowledge of the translational process. Furthermore, some antibiotics work by disabling the ribosome in bacterial cells, but not in human cells, allowing us to fight infections without harming ourselves. A complete understanding of the ribosome and how it differs between bacteria and humans was necessary to obtain our current understanding of these drugs.

As mentioned, deregulation in the processes pertaining to the central dogma frequently lead to disease. Deregulation in translation and the production of ribosomes specifically has been associated with cancer and an increasing number of inherited diseases. Given that one of the characteristics of cancer is uncontrolled growth, and that proteins are necessary for growth, it makes sense that the ribosome is associated with cancer. In fact, early experiments have shown that growth rate is correlated with ribosome content in *Escherichia coli* (Maaløe & Kjeldgaard, 1966). In *Drosophila melanogaster* (common fruit fly), the *minute* mutants (characterized by thin bristles, slow development, reduced viability, rough eyes and small body size) were found to result from mutations in genes encoding ribosomal proteins (Kongsuwan et al., 1985; Marygold et al., 2007). In *Saccharomyces cerevisiae* (brewer's yeast), growth has been shown to be correlated with cell division (Johnston et al., 1977; Jorgensen et al., 2002). These early studies have laid the groundwork for the current understanding of the connection between ribosomes and cell division.

Protein folding & structure

Ribosome biogenesis and translation are needed to produce proteins, the biomolecules that carry out the myriad of functions required for life. After having been created by the ribosome, the linear amino acid chain composed of the 20 commonly used amino acids (table 1) (also known as a polypeptide chain) will undergo several modifications before folding into a fully functioning protein. Correct protein folding is a crucial component of protein synthesis since the structure of a protein will determine its function. Misfolded proteins have been found to be associated with a number of diseases including Alzheimer's and Parkinson's disease (Chiti & Dobson, 2006). Furthermore, the process of protein synthesis is tightly regulated, and perturbations can also lead to disease. A protein contains four types of structures: primary, secondary, tertiary and quaternary structures (reviewed in Tymoczko, 2011).

Table 1. Amino acids and their commonly used abbreviations.

Amino acid	Three letter abbreviation	One letter abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Primary structure

The primary structure of a protein refers to the linear sequence of amino acids in a polypeptide chain (fig. 2). For example, the mature *Homo sapiens* insulin protein is 51 amino acids in length. It is made up of two main components, known as the A and B chain, which are linked together. The corresponding primary structure of the A chain of insulin can be denoted as follows: GIVEQCCTSICSLYQLENYCN. It is important to note that the insulin protein undergoes a number of modifications not mentioned here, as they are not relevant for this project.

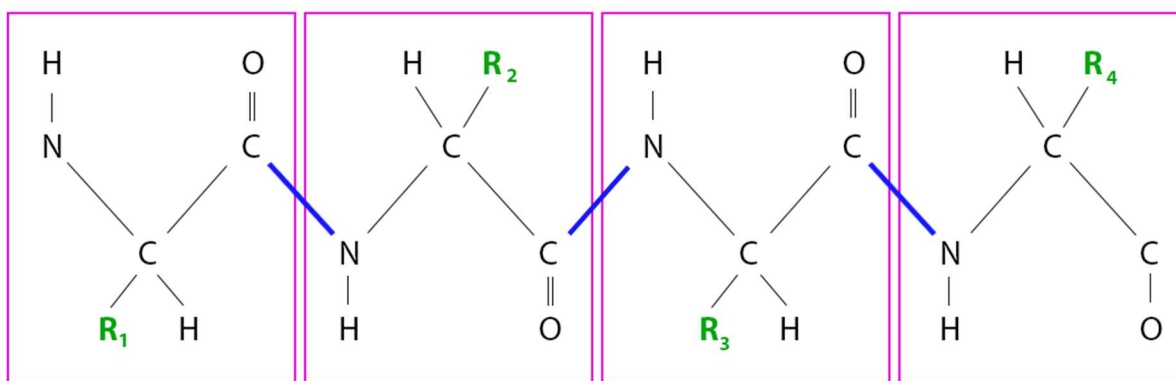


Figure 2. Schematic of a polypeptide chain. Each box depicts the basic chemical structure of an amino acid with the amino group (NH) on the left of the central carbon and the carbonyl group (CO) on the right of the central carbon. Peptide bonds between amino acids are highlighted in blue. This basic structure is the same across all amino acids, the difference between them is the chemical makeup of the side chain (also referred to as the R group), depicted in green.

Secondary structure

The secondary structure of a protein refers to the basic structures created through interactions between the backbones of the different amino acids. These structures can be α -helices, β -pleated sheets, turns or loops (fig. 3) and they are held together by hydrogen bonds between the carboxyl and amino termini of the different amino acids (fig. 2).

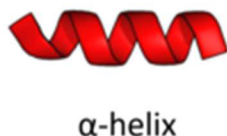
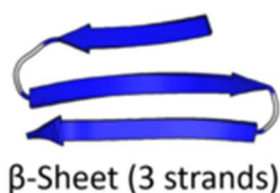


Figure 3. Ribbon representations of the two most common secondary structures: β -sheets (left) and α -helices (right).
Image from reference 2.

α -helices

Alpha helices are secondary structures characterized by the amino acid residues coiling around a central axis. Each amino acid is related to the other by a rise of $1,5\text{\AA}$ (\AA ngström, a unit of length equivalent to 0.1 nanometers) along the helix axis and a turn of 100 degrees, resulting in 3,6 amino acids per turn. It follows that amino acids that are four residues apart from each other in sequence are on top of each other in the helix, while amino acids that are two residues apart from each other are on opposite ends of the helix and therefore unlikely to touch each other (fig. 4). Helices are also said to have pitch, referring to one complete turn along the helix axis. Pitch can be calculated as the rise multiplied by the number of residues per turn (i.e. $1,5\text{\AA} \times 3,6 = 5,4\text{\AA}$). Finally, alpha helices also have a screw sense, which can be right-handed (clockwise) or left-handed (anti-clockwise) representing the direction in which the helix coils. All α -helices in proteins are right-handed as this conformation results in fewer steric clashes.

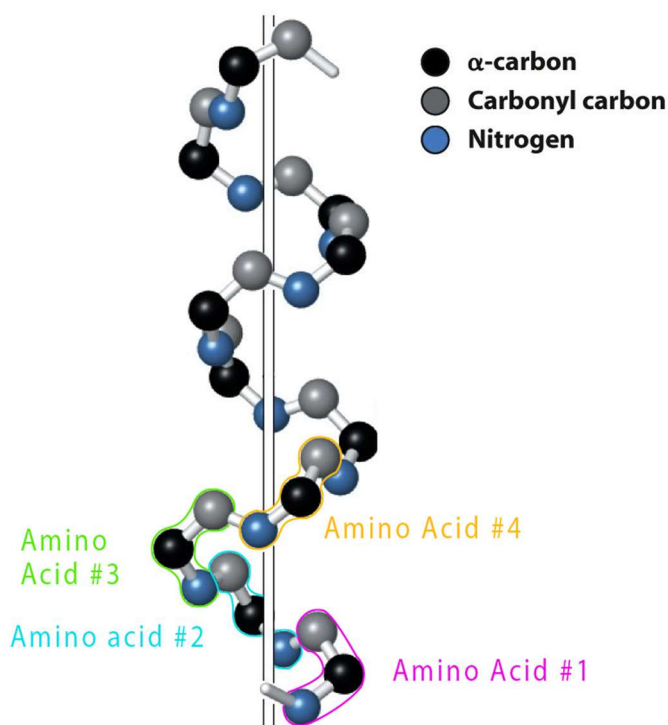


Figure 4. α -helix structure on the atomic level. An α -helix is shown as it wraps around its central axis (depicted as the two parallel black lines). Nitrogen atoms are shown in blue, the carbonyl carbon in grey and the α -carbon is shown in black. The atoms corresponding to the first four amino acids on the bottom of the structure are grouped by colour. Amino acids that are two residues apart in primary structure (e.g. amino acid #1 and #3) are furthest apart in three dimensional space and thus unlikely to interact, while amino acids four residues apart are directly on top of each other (e.g. amino acid #1 and #4). Image adapted from reference 3.

Certain amino acids tend to disrupt the formation of α -helices due to their chemical structure. Valine, threonine and isoleucine tend to disrupt their formation due to steric clashes. Serine, aspartate and asparagine tend to disrupt α -helices because their side chains contain hydrogen bond donors and acceptors that lie in close proximity to the main chain, resulting in competition for the main chain NH and CO group, thereby destabilizing the helix (table 2).

β -sheets

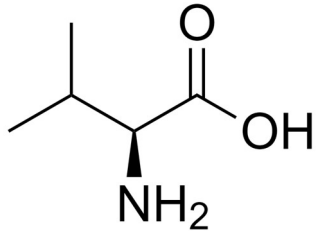
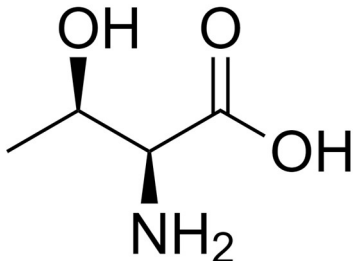
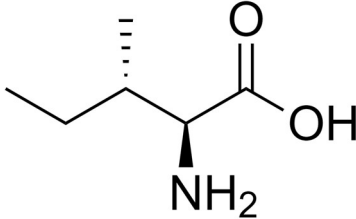
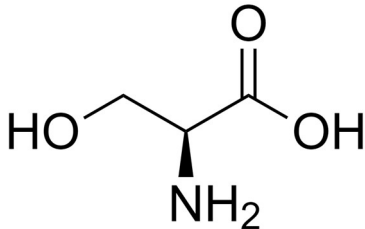
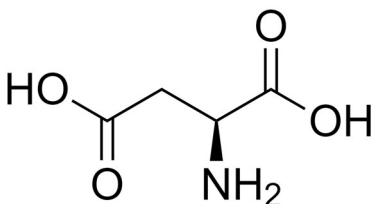
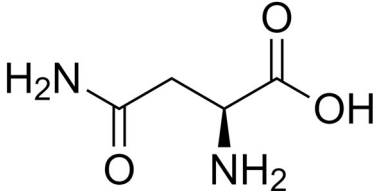
β -sheets are made up of two or more polypeptide chains known as β -strands. Rather than being coiled like α -helices, β -strands are fully extended. The distance between two amino acid residues is 3.5\AA , all of which are organized in a trans conformation to avoid steric clashes (fig. 4). β -sheets are formed by linking two or more β -strands lying next to each other through hydrogen bonds. In the case of an antiparallel β -sheet the β -strands run in opposite directions, while in the case of a parallel β -sheet, the β -strands run in the same direction. In contrast with α -helices, where all the residues are close to each other in the primary structure, β -sheets can be composed of residues that are far away from each other in the primary structure.

Loops and turns

Most proteins are made up of a combination of different secondary structures and take on globular architectures, which require strand reversals. These are accomplished by reverse turns and loops. The majority of these structures will lie on the surface of the protein, functionally they therefore often participate in interactions with other proteins and the environment. Loops and

turns exposed to an aqueous solution usually contain amino acid residues with hydrophilic R groups. For example, in insulin, chain A folds into two α -helices, between which a small β -strand is situated. Chain B on the other hand is made up of a larger α -helix.

Table 2. The chemical structure of common α -helix disrupting amino acids.

α -helix disruptors due to steric clashes		
Valine	Threonine	Isoleucine
		
α -helix disruptors due to hydrogen bonding		
Serine	Aspartate	Asparagine
		

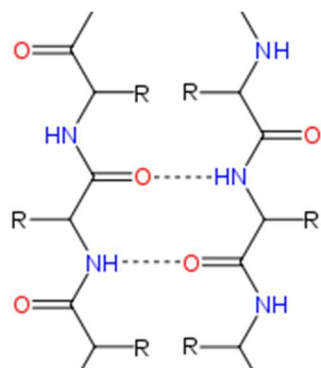


Figure 5. Molecular depiction of an antiparallel β -sheet. Oxygen atoms are shown in red and amino groups are shown in blue. Hydrogen bonds between the oxygen atom on one strand, and the hydrogen atom of the amino group on the other strand, are depicted as black dashes. Adjacent R groups on both strands are in a trans conformation to reduce steric hindrance, meaning that they alternate the side of the backbone they are on. The distance between two adjacent R groups is 3.5Å. Image adapted from reference 4.

Tertiary structure

The tertiary structure is the result of interactions between the R groups of the peptide chain of a protein and it determines its three dimensional structure. Tertiary structures involve the

arrangement of amino acids that are far apart from each other in sequence, as well as the pattern of disulfide bonds between the sulphur containing cysteine residues. In the case of insulin, the tertiary structure will be composed of chain A and chain B being linked together by a total of six bonds known as disulfide bonds.

Quaternary structure

Finally, the quaternary structure of a protein refers to the arrangement of subunits and their interactions. The types of forces involved are hydrogen bonds, ionic bonds and Van der Waal interactions. The quaternary structure of insulin consists of six insulin molecules (as described in the tertiary structure) that group together in the form of a doughnut shape.

The process of folding

Once formed, the polypeptide chain needs to fold by taking on its primary, secondary, tertiary and quaternary structure. Proteins fold by progressively stabilizing intermediate states, rather than randomly going through every possible conformation, a process that would take 1.6×10^{27} years for a 100 residue protein (assuming that each residue can take on three different configurations and the conversion of one structure into another takes 10^{-13} seconds) (Levinthal, 1968). One can think of an unfolded protein as being in a high energy state (very unstable) and a correctly folded protein as being in a low energy state (stable). Protein folding is the process of going from a high energy state to a low energy state. By progressively moving to lower energy states, a polypeptide chain retains partially correct intermediate steps that are slightly more stable than unfolded regions. Each correctly folded residue will contribute ~ 0.42 kJ/mol of energy to maintain a fold. In contrast, the amount of thermal energy at room temperature is 2.5 kJ/mol, this allows partially correct conformations to be destabilized easily.

Protein folding is driven by the hydrophobic effect: a polypeptide chains folds such that the hydrophobic residues are buried within the protein to avoid the contact with water, while polar and charged residues remain on the surface of the protein. Furthermore, unpaired amino and carboxyl groups prefer an aqueous environment to a non-polar environment due to their ability to hydrogen bond. In order to bury these residues within the protein it is necessary to pair them via hydrogen bonding. This is usually accomplished with β -sheets and α -helices.

Some proteins, or parts thereof, may be intrinsically unstructured and only assume a defined structure upon interaction with other proteins or other molecular structures. They can take on a number of states from being fully unstructured to being partially unstructured. These types of proteins appear to be particularly important in signalling and regulatory pathways, as they have the potential to bind to multiple partners, frequently by taking on a different conformation (Wright & Dyson, 2015). Intrinsically unstructured proteins are one of the four main types of proteins along with fibrous, globular and membrane proteins (Andreeva et al., 2013).

Protein domains & complexes

Protein domains are the basic structural units of a protein that can acquire their folded state independently (Schaeffer & Daggett, 2010), they are usually 50-350 amino acids in length and are made up of combinations of α -helices and β -sheets that pack together forming globular

units. Small proteins may only contain a single domain, while larger proteins are made up of a number of domains linked by open lengths of peptide chain. Protein complexes are formed by smaller subunits combining into one larger structure. These subunits are bound to each other by weak non-covalent interactions (Alberts et al., 1994). When studying proteins, the entire protein or individual domains of it can be studied. Moreover, across evolution, evidence of domain swaps between proteins and conservation of sequences can be used to infer domain boundaries. Similarly, the evolution of domains and their conservation can be used to infer function of their resident proteins.

Studying proteins

There are several methods that can be used to study proteins in general, these include methods used in both wet and dry labs. Given that the structure of a protein determines its function, determining its structure is a common approach taken when analyzing a novel protein. This can be done using X-ray crystallography, nuclear magnetic resonance (NMR) or cryo-electron microscopy (cryo-EM)¹. One of the benefits of NMR is the fact that it allows researchers to study how a protein moves in solution, thus giving insight into how it might behave *in vivo*. Depending on the overall size of a structure, some of these methods also allow for the structural determination of a protein and one or multiple of its interactors.

Another way to study disease associated proteins is to study the effect mutations or entire gene deletions have in model organisms. For example, experiments could examine the effect specific mutations have on an organism and if they are embryonically lethal, or, they could measure the effect they have on the location of certain proteins within the cell. Another commonly done experiment, usually in yeast, involves the creation of double mutants, which allow the researchers to determine the effect deletions in a gene of interest have when they are combined with deletions in other genes. Reproducing similar experiments in different model organisms are also of value to confirm that findings hold across different species, or to uncover differences in the way mechanisms work across life.

Finally, comparative evolutionary studies can also be conducted to study proteins using computational approaches. For the purposes of this research there are two main approaches: a population genetics approach and a evolutionary genomics approach. As the name suggests, the former deals with genetic variations across a population, and requires sequence data from a large number of individuals of the same species. Furthermore, population genetics focuses on how the dynamics within a population affect population structure and genetic diversity.

Evolutionary genomic approaches examine variations of a protein across the tree of life. The most basic analysis of this type is to identify and compare species that have a copy of the gene

¹ These methods map the location of the atoms in the protein, i.e. its structure, either by measuring the diffraction pattern of x-ray or electron beams colliding with the atoms of the protein (crystallography and cryo-EM, respectively) or measuring the energy released due to changes in quantum spin within the atoms of the protein upon the application and removal of a strong magnetic field (NMR).

of interest with those that are missing the gene entirely. The process a gene is involved with is likely conserved in some way among the species that share the protein. If some species are missing the gene entirely, this could suggest that the process under study does not happen in these species, the process differs in some way and does not require the protein under study, or a different protein takes on its function. When characterizing a novel gene, identifying species that share a copy is thus a good starting point to generate hypotheses about its function. If information on the process the protein is involved with is available, knowing what species have a copy of the gene will also allow for comparisons of the differences in this process within these species, thus further refining our understanding of its hypothesized function.

Sequences of the same protein from different species can also give insight on the importance of individual residues for its function. Given that the sequence of amino acids of a protein determines its structure and therefore function, sequence variation across species can be used to infer if its function is likely to be conserved, or if there are differences between species. This is done by creating multiple sequence alignments of the protein sequences.

Multiple sequence alignments

By creating multiple sequence alignments, algorithms attempt to group the same residues of a protein in the same column to allow for a comparison between sequences. This concept is best illustrated by comparing it to a 'multiple word alignment' of the word coffee (fig. 6). Here the algorithm attempts to group identical or similar letters in the same column. Some sequences may have deletions or insertions, thus increasing or decreasing their length. Alignment algorithms need to take this into account by creating gaps in the alignment, e.g. letters three and five ('f' and 'e') in Dutch, German and English in the word example could show an insertion in these languages, or a deletion of these letters in the other languages. Alignment algorithms take this into account by creating gaps (denoted by dashes).

Dutch:	K	o	f	f	i	e
German:	K	a	f	f	e	e
Spanish:	C	a	-	f	-	é
Catalan:	C	a	-	f	-	è
English:	C	o	f	f	e	e
Portuguese:	C	a	-	f	-	é
Greek:	κ	α	-	φ	-	έ

Figure 6. A 'multiple word alignment' of the word coffee in different languages. Letters are aligned in such a way that the same or similar letters are in the same column. In this example all permutations of the letter 'e' and 'f' are assumed to have similar characteristics, the letters 'c' and 'k' are also assumed to have the same characteristics, while the letters 'o' and 'a' are assumed to have different characteristics. The letters are colour coded according to these characteristics. Columns coloured with a single colour thus show conservation across all languages. The same type of analysis can be done with genomic sequences.

Multiple sequence alignments can shed light on the importance of individual residues as well as larger sections of the protein. If a residue or section of the protein is conserved across all of the species studied, this suggests that it takes on an important role in the function of the protein.

Similarly, sections of a protein that do not show any conservation are likely not that important for its function, since any residue is tolerated. Multiple sequence alignments can shed further light on the function of a protein if it contains an extra section or specific changes in certain species. If these sections are conserved, this might suggest that they have some additional functionality not found in the species that lack the additional section. Similarly, if certain species lack a section of the protein that is known to have an important function in other organisms, this raises the question on how the process differs between species. One example of this type of analysis has been conducted using the *FOXP2* gene that is known to be involved in the acquisition of speech and language in humans. Based on a multiple sequence alignment, Enard et al. (2002) discovered that the FOXP2 protein acquired two amino acid changes specific to the human homolog in an otherwise conserved region. This led to the hypothesis that FOXP2 might have been involved in the process of language acquisition in humans. Besides their usage in comparative protein evolution, multiple sequence alignments are also necessary for many methods of phylogenetic tree reconstruction.

Just as species evolve and diverge over time, protein sequences also evolve and diverge. This allows us to create phylogenetic protein trees, also known as gene trees, that depict how a protein evolved. For the purposes of this type of analysis, we are interested in genes from different species that have the same common ancestor. These are also known as homologs. There are two main ways homologs can arise: through speciation or through gene duplication. Genes that arise through speciation are known as orthologs, while genes that arise through gene duplications are known as paralogs (Fitch, 1970).

The advent of high-throughput genetic sequencing and its continuously falling cost has significantly increased the number of available sequences in public databases that can be used by scientists to conduct multiple sequence alignments and phylogenetic analyses. In 2003 for example, sequences from only around 2,000 organisms were available. Today, sequences for almost 80,000 organisms are available, representing a 40x increase in only 15 years (RefSeq growth statistics, (n.d.)). This increase in available sequences allows researchers to make comparisons previously not possible, making a renewed visit of early studies a potentially worthwhile endeavour.

Shwachman-Diamond syndrome

As mentioned, translation is one of the fundamental processes of life and can be studied in humans by studying disease. One of the diseases that is caused by the impaired production of ribosomes, specifically impaired subunit joining, is the ribosomopathy Shwachman-Diamond syndrome (SDS, OMIM 260400), named after two of the doctors involved in its discovery (Bodian et al., 1964; Shwachman et al., 1964). SDS is an autosomal recessive disorder that affects around one in 76,000 people (Goobie et al., 2001). Its symptoms include pancreatic insufficiency, leading to digestive issues, skeletal abnormalities, short stature, cognitive impairment, structural brain alterations and a cumulative risk of leukemia of about about 36% by the age of 30 (Donadieu et al., 2005; Ginzberg et al., 1999; Kerr et al., 2010; Mack et al., 1996;

Toiviainen-Salo et al., 2008). These symptoms make SDS an important model to aid our understanding of the genetic determinants involved in the multi-step progression to leukemia, as well as the fundamental process of ribosome biogenesis itself.

Around 90% of SDS cases can be attributed to biallelic mutations in the Shwachman-Bodian-Diamond Syndrome (*SBDS*) gene, which is conserved across both eukaryotes and archaea (Boocock et al., 2003; Boocock et al., 2006; Dror et al., 2011). *SBDS* is located on the long arm of chromosome 7, it is made up of five exons spanning 7.9 kb and encodes a 1.7 kb transcript that is translated to create a 250 amino acid protein (Boocock et al., 2003). The region around *SBDS* is locally duplicated and contains a pseudogene copy (a non-functional version of the gene) of *SBDS*, known as *SBDSP*. The transcript of *SBDSP* is 97% identical to *SBDS* and contains nucleotide changes and deletions that disrupts its protein coding potential. Notably, the *SBDS* homolog in the highly studied plant model organism *Arabidopsis thaliana* has an extended C-terminal domain containing a predicted zinc-finger domain fusion (Boocock et al., 2003).

Disease causing SBDS mutations

There are a number of known disease associated mutation in *SBDS* (table 3). The two most common mutations both cause premature protein truncations (i.e. they result in shorter, non-functional proteins due to prematurely stopped translation). The first is a dinucleotide change from TA to CT at positions 183-184 (denoted as 183-184TA>CT) resulting in an in frame stop codon (denoted as K62X). The second mutation is a single nucleotide change from T to C, two nucleotides within intron 2 at position 258+2 (denoted as 258+2T>C) disrupting a splice site, resulting in a 8 base pair deletion. This deletion will end up causing a premature truncation of the protein via frameshift (denoted as 84Cfs3) (Boocock et al., 2003).

Multiple sequence alignments of the *SBDS* gene of affected individuals, controls and other sequences from GenBank showed that both of these mutations naturally occur in *SBDSP*. Based on this observation, it became clear that both of these mutations arose from recombination-based gene conversion between *SBDS* and *SBDSP* (Boocock et al., 2003). Gene conversion was found in 89% of unrelated individuals that had mutations in *SBDS*, with 60% of them carrying two converted alleles. Gene conversions account for 74.4% of alleles associated with SDS. All other known disease-causing mutations are deletions, insertions or point mutations (table 3).

Table 3. List of known malignant mutations in *SBDS*.

Nucleotide sequence change	Predicted protein consequence	Second allele	Number of families (F) /Individuals (I)	Reference
Exon 1				
c. 13del	Thr5Profs*8	None found ¹	1 (I)	Donadieu et al., 2012
c. 17 C>T	Pro6Leu	c. 258+2T>C	1 (I)	Donadieu et al., 2012
c. 24C>A	Asn8Lys	c. 258+2T>C	1 (I)	Boocock et al., 2003
c. 56G>A	Arg19Gln	c. 258+2T>C	1 (I)	Shammas et al., 2003
c. 79T>C ²	Phe27Leu	c. 183-184delInsCT ²	1 (I)	Nishimura et al., 2007
c. 93C>G	Cys31Trp	c. 258+2T>C	1 (F)	Shammas et al., 2003
c. 95A>G	Tyr32Cys	c. 258+2T>C	2 (I)	Nicolis et al., 2005; Rosendahl et al., 2006
c. 96-97insA	Asn34Lysfs*16	c. 258+2T>C	2 (I)	Boocock et al., 2003; Nakashima et al., 2004
c. 97A>G	Lys33Glu	c. 258+2T>C	1 (F)	Shammas et al., 2005
c. 98A>C ³	Lys33Thr	None found ¹	1 (I)	Shah et al., 2009
c. 101A>T	Asn34Ile	c. 258+2T>C	1 (I) & 1 (F)	Nicolis et al., 2005; Shammas et al., 2005
c. 101A>G	Ans34Ser	183-184delInsCT (Phase unknown)	1 (I)	Newman et al., 2009
c. 107del	Val36Alafs*23	c. 258+2T>C	1 (I)	Maserati et al., 2006
c. 120del (c.119del)	Ser41Alafs*18	c. 258+2T>C	6 (I)	Boocock et al., 2003; Mäkitie et al., 2004; Austin et al., 2005
c. 123del	Ser41Argfs*18	c. 258+2T>C	1 (F)	Shammas et al., 2005
c. 127G>T ⁴	Val43Leu	None found ¹	1 (I)	Karow et al., 2010
Intron 1				
c. 129-2A>G	~~	None found ¹	1 (I)	Donadieu et al., 2012

c. 129-1G>A	Glu44fs*1	None found ¹	1 (I)	Donadieu et al., 2012
c. 129-71_140del83	~~	c. 258+2T>C	1 (I)	Maserati et al., 2006
Exon 2				
c. 129-?_258+?	Exon 2 deletion	c. 258+2T>C	1 (I)	Donadieu et al., 2012
c. 131A>G	Glu44Gly	c. 258+2T>C	3 (I)	Boocock et al., 2003; Mäkitie et al., 2004; Tsangaris et al., 2012
c. 164C>A	Ser55*	None found ¹	1 (I)	Donadieu et al., 2012
c. 171T>A	Phe57Leu	c. 258+2T>C	1 (I)	Donadieu et al., 2012
c. 183-184delInsCT*	Lys62*	c. 258+2T>C	90 (I) & 135 (F)	Boocock et al., 2003; Donadieu et al., 2012; Kuijpers et al., 2005; Toiviainen-Salo et al., 2008; Taneichi et al., 2006; Mäkitie et al., 2004; Hashmi et al., 2011; Xia et al., 2009; Tsangaris et al., 2012; Woloszynek et al., 2004; Austin et al., 2005; Keogh et al., 2012; Church, 2006; Kawakami et al., 2005; Mellink et al., 2004; Booij et al., 2013
c. 183-184delInsCT*	Lys62*	c. 258+2T>C (Phase unknown) ²	1 (I)	Nishimura et al., 2004
c.[183-184delInsCT; 258+2T>C]*	Lys62*	c. 258+2T>C	19 (I) & 14 (F)	Boocock et al., 2003; Donadieu et al., 2012; Nicolis et al., 2005; Maserati et al., 2006; Mäkitie et al., 2004; Hashmi et al., 2011; Tsangaris et al., 2012
c. 199A>G	Lys67Glu	c. 258+2T>C	1 (I)	Boocock et al., 2003
c. 212T>C	Leu71Pro	c. 258+2T>C	1 (I) & 1(F)	Shammas et al., 2005
c. 250T>C	Cys84Arg	c. 258+2T>C	1 (I)	Kuijpers et al., 2005
c. 258+1G>C	Ile87Leufs*15	c. 258+2T>C	3 (I)	Boocock et al., 2003; Woloszynek et al., 2004
c. 258+2T>C ⁵	Ile87Alafs*15 / (p.Cys84Tyrfs*4)	c. 258+2T>C	5 (I) & 7 (F)	Boocock et al., 2003; Donadieu et al., 2012; Nicolis et al., 2005; Maserati et al., 2006; Tsangaris et al., 2011; Austin et al., 2005
c. 258+2T>C	Ile87Alafs*15 / (p.Cys84Tyrfs*4)	c. 259-124G>A ⁷	1 (I)	Toiviainen-Salo et al., 2008

c. 258+2T>C	Ile87Alafs*15 / (p.Cys84Tyrfs*4)	None found ^{1,8}	8 (I)	Kuijpers et al., 2005; Maserati et al., 2006; Toiviainen-Salo et al., 2008; Karow et al., 2010; Mäkitie et al., 2004; Xia et al., 2009; Kawakami et al., 2005; Khan et al., 2008
Intron 2				
c. 259-1G>A	Ile87Valfs*15	c. 258+2T>C	1 (I)	Taneichi et al., 2006
Exon 3				
c. 258+374_459+250del	Ile87_Gln153del	c. 258+2T>C	1 (I)	Costa et al., 2007
C. 260T>C	Ile87Thr	c. 258+2T>C	1 (I)	Mäkitie et al., 2004
c. 260T>G	Ile87Ser	c. 258+2T>C	2 (I)	Boocock et al., 2003; Tsangaris et al., 2012
c. 279_284del	Gln94_Val95del	c. 258+2T>C	1 (F)	Shammas et al., 2005
c. 291_293delIns AGTTCAAGTATC	Asp97- Lys98delInsEVQVS	c. 258+2T>C	1 (I)	Boocock et al., 2003
c. 297_300ddelAAGA (c. 292_295delAAAG)*	Glu99Aspfs*21	c. 258+2T>C	5 (I) & 3 (F)	Shammas et al., 2005; Kuijpers et al., 2005; Kawakami et al., 2005; Booij et al., 2013
c. 307_308del	Gln103Thrfs*6	c. [258+2T>C; 201A>G]	1 (I)	Nicolis et al., 2005
c. 354A>C	Lys118Asn	c. 183-184delInsCT	1 (F)	Shammas et al., 2005
c. 355T>C	Cys119Arg	c. 258+2T>C	1 (I)	Donadieu et al., 2012
c. 356G>A	Cys119Tyr	c. 258+2T>C	3 (I)	Donadieu et al., 2012
c. 362A>C ⁹	Ans121Thr	c. 523C>T	1 (I)	Erdos et al., 2006
c. 377G>C	Arg126Thr	c. 258+2T>C	2 (I)	Boocock et al., 2003
c. 385A>G	Thr129Ala	None found ¹	1 (I)	Donadieu et al., 2012
c. 388G>T	Val130Leu	c. 183-184delInsCT	1 (I)	Hashmi et al., 2011
c. 428C>G	Ser143Trp	c. 258+2T>C	1 (I)	Taneichi et al., 2006

c. [428C>T; 443A>G]*	[Ser143Leu; Lys148Arg]	c. 258+2T>C	1 (F)	Shammas et al., 2005
c. 443A>C	Lys148Thr	c. 183-184delInsCT	1 (I)	Ball et al., 2009
c. 453A>C	Lys151Ans	None found ¹	1 (I)	Donadieu et al., 2012
c. 458A>G	Gln153Arg	c. 258+2T>C	1 (F)	Shammas et al., 2005
~~	Q94* ¹⁰	None found ¹	?	Xia et al., 2009
Intron 3				
c. 460-1G>A	Ala154Valfs*18	c. 258+2T>C	1 (I) + 1 (F)	Shammas et al., 2005; Austin et al., 2005
Exon 4				
c. 461C>T	Ala154Val	c. 461C>T	1 (I)	Donadieu et al., 2012
c. 505C>T	Arg169Cys	c. 258+2T>C	3 (I)	Shammas et al., 2005; Woloszynek et al., 2004; Austin et al., 2005
c. 506G>T	Arg169Leu	c. 258+2T>C	1 (I) & 1 (F)	Shammas et al., 2005; Donadie et al., 2012
c. 506G>A	Arg169His	c. 258+2T>C	1 (I)	Rommens lab, unpublished.
c. 523C>T ⁹	Arg175Trp	c. 362A>C	1 (I)	Erdos et al., 2006
Intron 4				
c. 624+1G>C	Val209Leufs*18	c. 258+2T>C	2 (I) & 1 (F)	Shammas et al., 2005; Nicolis et al., 2005
c. 624+1G>A	Val209Ilefs*18	c. 258+2T>C	1 (I)	Tsangaris et al., 2012
Exon 5				
c. 652C>T	Arg218*	c. 258+2T>C	2 (I)	Woloszynek et al., 2004
c. 653G>A	Arg218Gln	c. 258+2T>C	1 (I)	Donadieu et al., 2012

¹ Exon sequencing was performed

² Associated with Spondylometaphysial dysplasia (SMD) resembling SMD Sedaghatian type

³ Associated with primary immunodeficiency (SDS diagnosis was considered)

⁴ Associated with refractory cytopenia, authors suggested c.127G>T may be a benign variant

⁵ The exome variant server lists 32/8600 alleles; 0,37% of European American population

⁶ Description based on experimental evidence

⁷ Variant may be benign

⁸ Allele identified in individuals with SDS (Kuijpers et al., 2005; Toiviainen-Salo et al., 2008; Mäkitie et al., 2004; Kawakami et al., 2005) refractory cytopenia (Karow et al., 2010) severe congenital cytopenia (Xia et al., 2009) variable immunodeficiency (Khan et al., 2008) and aplastic anemia (Maserati et al., 2006).

⁹ Individual was compound heterozygote with another mutation

¹⁰ Associated with severe congenital cytopenia

*Mutations that arise from gene conversions are in bold font

Table 4. Benign mutations in *SBDS*.

Nucleotide sequence change	Predicted protein consequence	Reference	Allele frequency	
			SDS individuals/families	Population study
Exon 4				
c. 501A>G	Ile167Met	Nakashima et al., 2004	NA	1/140 alleles (0.71%)
c. 572_573insA	Pro192Alafs*9	Exome variant server	NA	24/8254 alleles (0.29%)
Exon 5				
653T>C	Ile212Thr	Boocock et al., 2003; Nicolis et al., 2005; Karow et al., 2010; Exome variant server	2/316 alleles (0.63%) (Boocock et al., 2003) 5/30 alleles (16.5%) (Nicolis et al., 2005)	334/8600 alleles (3.88%) (Exome variant server)

The three dimensional structure of SBDS

In order to better understand the molecular consequences of mutations in *SBDS*, early studies set about solving the three dimensional crystal structure of the *Archaeoglobus fulgidus* *SBDS* homolog (AfSBDS) (Savchenko et al., 2005; Shammass et al., 2005). They concluded that AfSBDS has a three domain architecture (fig. 7). The first domain spans residues D5-I87, the second domain spans T88-F161 and the third domain spans E162-G234.

The first domain was found to contain four β -strands forming a three stranded antiparallel β -sheet, and four α -helices (fig. 7). It was also found to be a structural structural homolog with a single domain protein in *S. cerevisiae*, known as *Yhr087wp*, despite a sequence identity of only 15.3% (Savchenko et al., 2005). This was not surprising since proteins of the same structure can sometimes have very different sequences. The distribution of *Yhr087wp* was found to be restricted to fungi. In light of the structural homology between AfSBDS and *Yhr087wp*, and the restricted distribution of the latter protein to fungi, domain I was named the FYSH domain (Fungi, Yhr087wp, Shwachman) (Shammass et al., 2005).

The central domain (domain II) consists of three α -helices ($\alpha 5$ - $\alpha 7$). Helices $\alpha 5$ and $\alpha 6$ were found to be connected by a conserved proline rich loop (fig. 7)

The C-terminal domain (domain III) is made up of two α -helices and four β -strands, forming a four-stranded antiparallel β -sheet. The two α -helices were determined to pack against the concave surface of the β -sheet (fig. 7, fig. 8). The $\beta\alpha\beta\beta\alpha\beta$ fold of this domain is typical of a ferredoxin-like fold. Despite no obvious homology between the sequences, the closest structural homolog to this domain was determined to be domain V of elongation factor 2 (*Ef2*) in *S. cerevisiae* (Savchenko et al., 2005). Both *Ef2* and its bacterial counterpart EF-G (elongation factor G, formerly known as translocase) are GTPases (a family of enzymes that can bind and hydrolyze guanosine triphosphate, GTP, a common substrate in many biochemical reactions). EF-G specifically catalyzes the translocation of the mRNA and tRNAs through the ribosome (Shoji et al., 2009).

Furthermore, the $\beta\alpha\beta\beta\alpha\beta$ fold is a common fold that shares structural homology with the RNA recognition motif (RRM), which gave us the first clue that SBDS might be binding to RNA. Additional evidence that SBDS is involved with RNA metabolism was provided by findings that a yeast SBDS homolog (*Ylr022c*) associated with ribosomal proteins, and other proteins associated with rRNA processing, as well as a computational study that predicted SBDS to function in rRNA processing (Savchenko et al., 2005). The latter was based on an analysis of protein-protein interactions and an identification of biologically relevant functional groups (Savchenko et al., 2005).

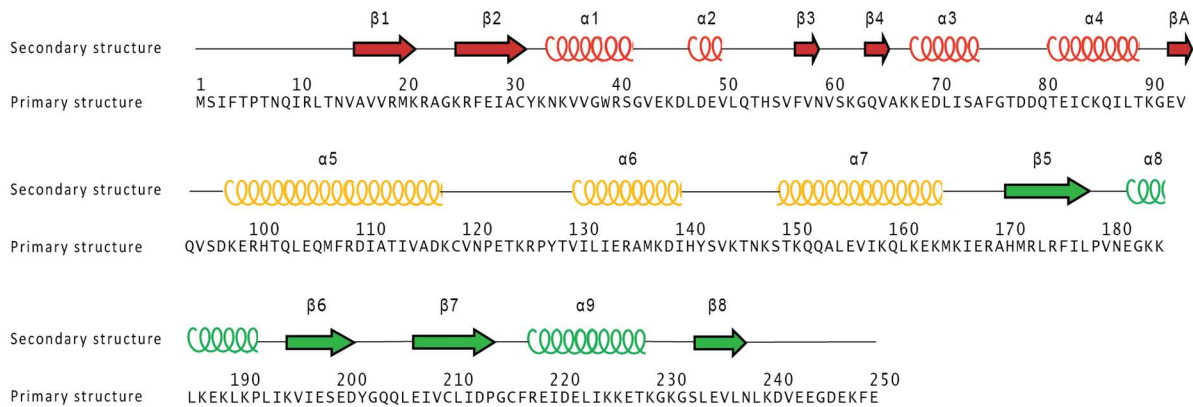


Figure 7. Schematic of the primary and secondary structure of the *H. sapiens* SBDS homolog. Structures coloured in red are part of the FYSH domain, structures coloured in yellow are part of the second domain and structures coloured in green are part of the third domain. β -sheets are shown as arrows while α -helices are shown as coils. Image adapted from Finch et al. (2011).

Based on the three dimensional structure of AfSBDS and SBDS, Shammas et al. (2005) and Finch et al. (2011) conducted an analysis of the impact of SDS-associated mutations on protein stability. They found that domains I and II are the most frequent target of SDS-associated mutations. Furthermore, Finch et al. (2011) classified the known mutations that affect protein stability or fold as class A mutations, and those that change surface epitopes without affecting overall stability or fold as class B mutations (fig. 8). Despite these recent studies, the exact impact of *SBDS* mutations on clinical SDS phenotypes are still unclear.

Shortly after the structure of AfSBDS was solved, the evolution and function of SBDS was investigated by several groups. Complementation assays in *S. cerevisiae* using the complete coding sequence or domain constructs of *SBDS* homologs from different species determined that SBDS functions in a species specific manner (Boocock et al. 2006). Furthermore, these experiments found the FYSH domain to be largely interchangeable among eukaryotic organisms, domain II was found to convey species specificity to protein function, and domain III was found to be largely dispensable for SBDS function. Additionally, evidence that SBDS functions in ribosome metabolism was provided by the localization of *SBDS* in a superoperon involved in RNA metabolism, as well as genetic and protein interactions that implicated a role in subunit joining (Krogan et al., 2006; Menne et al., 2007). These findings were further corroborated by subsequent studies in human cell lines (Burwick et al., 2012), as well as several model organisms including mice (Turlakis et al., 2012; Zhang et al., 2006), zebrafish (Provost et al., 2012), and *Dictyostelium* (Wong et al., 2011). Ultimately, these studies supported the current model that SBDS is an essential gene that functions in the late stages of ribosome subunit joining. Interestingly, there is some variability in the molecular signatures (e.g. ribosome phenotype and impact on translation) observed across species when SBDS is perturbed suggesting species specific roles. What specific changes in sequence are responsible for this species specificity remains an open question.

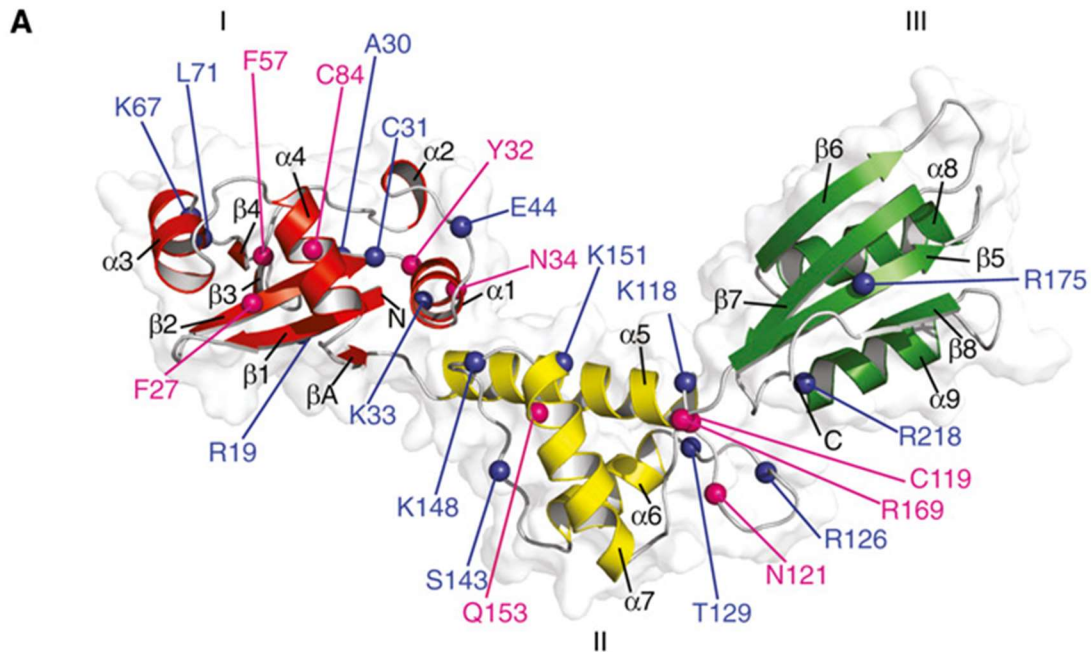


Figure 8. Three dimensional NMR based structure of the *H. sapiens* SBDS protein (Finch et al., 2011). The FYSH domain, second domain and third domain are shown in red, yellow and green respectively. Known SDS-associated mutations are mapped onto the structure, class A mutations that affect the stability of the protein are shown in pink, while class B mutations that affect surface epitopes are shown in blue.

In an attempt to gain a better understanding of its conservation across species, Boocock et al. (2006) attempted to find all *SBDS* homologs based on the available sequence data. Their searches and multiple sequence alignments identified homologs from a total of 159 organisms with broad conservation of the *SBDS* gene, representatives from all sequenced archaeal and eukaryotic genomes and all eukaryotic kingdoms were found. Only 228 bacterial genome sequences had been published at that time, and none of them appeared to contain *SBDS* homologs. Today over 50,000 bacterial reference sequences have been sequenced (RefSeq growth statistics, (n.d.)). A renewed search for *SBDS* homologs in bacteria could thus contradict these findings or confirm them with a greater level of confidence. Furthermore, only 18 *SBDS* homologs were found in archaea, limiting the comparisons that could be drawn between eukaryotic and archaeal *SBDS* and how they might differ. Residue Gly91 was found to be the only conserved residue across all eukaryotic and archaeal sequences. Given the increase in available sequences, it remains to be determined if Gly91 is invariant across all species. Finally, alignments based on additional sequences could be analyzed in light of recently identified structurally important sections of *SBDS* (Finch et al., 2011; Weis et al., 2015) and additional benign and malignant mutations that have recently been reported. Comparisons between different taxa could then uncover areas of *SBDS* that might differ in function between species, providing further insight into the function of this important disease gene. Here we set out to

update the phylogenetic analysis of SBDS in light of a ~40 fold increase in sequencing data. We confirmed the absence of *SBDS* in bacteria, created multiple sequence alignments with additional *SBDS* sequences, and analyzed the conservation of this gene in light of recently identified residues of structural and functional importance.

The Methods, Results and Discussion sections of this paper are under embargo pending manuscript preparation.

Image references

Image 1: a depiction of a translating ribosome. Adapted from www.shutterstock.com/image-vector/interaction-ribosome-mrna-process-initiation-translation-285444794

Image 2: Ribbon representations of the two most common secondary structures. From https://en.wikipedia.org/wiki/Protein_secondary_structure

Image 3: α -helix structure on the atomic level. Adapted from http://www.nslc.wustl.edu/courses/bio2960/labs/02Protein_Structure/PS2011.htm

Image 4: Molecular depiction of an antiparallel β -sheet. Adapted from: https://en.wikipedia.org/wiki/Beta_sheet

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. Molecular Biology of the Cell (Garland, New York, 1994). Google Scholar, 907-982.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2013). SCOP2 prototype: a new approach to protein structure mining. Nucleic acids research, 42(D1), D310-D314.

Austin KM, Leary RJ, Shimamura A. The Shwachman-Diamond SBDS protein localizes to the nucleolus. Blood. 2005 Aug 15;106(4):1253-8. Epub 2005 Apr 28.

Ball HL, Zhang B, Riches JJ, Gandhi R, Li J, Rommens JM, Myers JS. Shwachman-Bodian Diamond syndrome is a multi-functional protein implicated in cellular stress responses. Hum Mol Genet. 2009 Oct 1;18(19):3684-95. Epub 2009 Jul 14.

Basu, U., Si, K., Warner, J. R., & Maitra, U. (2001). The Saccharomyces cerevisiae TIF6 gene encoding translation initiation factor 6 is required for 60S ribosomal subunit biogenesis. Molecular and cellular biology, 21(5), 1453-1462.

Bernal, A., Ear, U., & Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Research, 29(1), 126-127.

Birkeland, N. K., Schönheit, P., Poghosyan, L., Fiebig, A., & Klenk, H. P. (2017). Complete genome sequence analysis of *Archaeoglobus fulgidus* strain 7324 (DSM 8774), a hyperthermophilic archaeal sulfate reducer from a North Sea oil field. *Standards in genomic sciences*, 12(1), 79.

BODIAN, M., SHELDON, W., & LIGHTWOOD, R. (1964). Congenital hypoplasia of the exocrine pancreas. *Acta paediatrica*, 53(3), 282-293.

Boocock, G. R. B., Marit, M. R., & Rommens, J. M. (2006). Phylogeny, sequence conservation, and functional complementation of the SBDS protein family. *Genomics*, 87(6), 758-771.

Boocock, G. R., Morrison, J. A., Popovic, M., Richards, N., Ellis, L., Durie, P. R., & Rommens, J. M. (2003). Mutations in SBDS are associated with Shwachman–Diamond syndrome. *Nature genetics*, 33(1), 97.

Booij J, Reneman L, Alders M, Kuijpers TW. Increase in central striatal dopamine transporters in patients with Shwachman-Diamond syndrome: additional evidence of a brain phenotype. *Am J Med Genet A*. 2013 Jan;161A(1):102-. Epub 2012 Dec 14

Brina, D., Miluzio, A., Ricciardi, S., & Biffo, S. (2015). eIF6 anti-association activity is required for ribosome biogenesis, translational control and tumor progression. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1849(7), 830-835.

Burwick, N., Coats, S. A., Nakamura, T., & Shimamura, A. (2012). Impaired ribosomal subunit association in Shwachman-Diamond syndrome. *Blood*, blood-2012.

Calado RT, Graf SA, Wilkerson KL, Kajigaya S, Ancliff PJ, Dror Y, Chanock SJ, Lansdorp PM, Young NS. Mutations in the SBDS gene in acquired aplastic anemia. *Blood*. 2007 Aug 15;110(4):1141-6. Epub 2007 May 3.

Chamberlain, S. A., & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*, 2.

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., O'Donnell, J., Oksanen, J., Greshake Tzovaras, B., Marchand, P., & Tran, V. (2018) taxize: A taxonomic toolbelt for R. R, rOpenSci. Retrieved from <https://github.com/ropensci/taxize> (Original work published 2011)

Chiti, F., & Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75, 333-366.

Church JA. A pediatric genetic disorder diagnosed in adulthood. *PLoS Med*. 2006 Jan;3(1):e15. Epub 2006 Jan 31.

Cormen, T. H., Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). Introduction To Algorithms. MIT Press.

Costa E, Duque F, Oliveira J, Garcia P, Gonçalves I, Diogo L, Santos R. Identification of a novel AluSx-mediated deletion of exon 3 in the SBDS gene in a patient with Shwachman-Diamond syndrome. *Blood Cells Mol Dis*. 2007 Jul-Aug;39(1):96-101. Epub 2007 Mar 21.

Dhanraj, S., Matveev, A., Li, H., Lauhasurayotin, S., Jardine, L., Cada, M., ... & Vincent, A. (2017). Biallelic mutations in DNAJC21 cause Shwachman-Diamond syndrome. *Blood*, 129(11), 1557-1562.

Donadieu, J., Fenneteau, O., Beaupain, B., Beauvils, S., Bellanger, F., Mahlaoui, N., ... & Perot, C. (2012). Classification and risk factors of hematological complications in a French national cohort of 102 patients with Shwachman-Diamond syndrome. *Haematologica*, haematol-2011.

Donadieu, J., Leblanc, T., Meunier, B. B., Barkaoui, M., Fenneteau, O., Bertrand, Y., ... & Bordigoni, P. (2005). Analysis of risk factors for myelodysplasias, leukemias and death from infection among patients with congenital neutropenia. Experience of the French Severe Chronic Neutropenia Study Group. *haematologica*, 90(1), 45-53.

Dror, Y., Donadieu, J., Koglmeyer, J., Dodge, J., Toiviainen-Salo, S., Makitie, O., ... & Cipolli, M. (2011). Draft consensus guidelines for diagnosis and treatment of Shwachman-Diamond syndrome. *Annals of the New York Academy of Sciences*, 1242(1), 40-55.

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic acids research*, 43(W1), W389-W394.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., ... & Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418(6900), 869.

Englbrecht, C. C., Schoof, H., & Böhm, S. (2004). Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. *BMC genomics*, 5(1), 39.

Erdos M, Alapi K, Balogh I, Oroszlán G, Rákóczi E, Sümegi J, Maródi L. Severe Shwachman-Diamond syndrome phenotype caused by compound heterozygous missense mutations in the SBDS gene. *Exp Hematol*. 2006 Nov;34(11):1517-21.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: <http://evs.gs.washington.edu/EVS/>) (Sep, 2013).

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2), 99-113.

- Gartmann, M., Blau, M., Armache, J. P., Mielke, T., Topf, M., & Beckmann, R. (2010). Mechanism of eIF6-mediated inhibition of ribosomal subunit joining. *Journal of Biological Chemistry*, jbc-C109.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, 31(13), 3784-3788.
- Ginzberg, H., Shin, J., Ellis, L., Morrison, J., Ip, W., Dror, Y., ... & Rommens, J. M. (1999). Shwachman syndrome: phenotypic manifestations of sibling sets and isolated cases in a large patient cohort are similar. *The Journal of pediatrics*, 135(1), 81-88.
- Goobie, S., Popovic, M., Morrison, J., Ellis, L., Ginzberg, H., Boocock, G. R., ... & Hudson, T. J. (2001). Shwachman-Diamond syndrome with exocrine pancreatic dysfunction and bone marrow failure maps to the centromeric region of chromosome 7. *The American Journal of Human Genetics*, 68(4), 1048-1054.
- Hashmi, S. K., Allen, C., Klaassen, R., Fernandez, C. V., Yanofsky, R., Shereck, E., ... & Samson, Y. (2011). Comparative analysis of Shwachman-Diamond syndrome to other inherited bone marrow failure syndromes and genotype–phenotype correlation. *Clinical genetics*, 79(5), 448-458.
- Johnston, G. C., Pringle, J. R., & Hartwell, L. H. (1977). Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*. *Experimental cell research*, 105(1), 79-98.
- Jorgensen, P., Nishikawa, J. L., Breitskreutz, B. J., & Tyers, M. (2002). Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297(5580), 395-400.
- Kapp, L. D., & Lorsch, J. R. (2004). The molecular mechanics of eukaryotic translation. *Annual review of biochemistry*, 73(1), 657-704.
- Karow, A., Flotho, C., Schneider, M., Fliegauf, M., & Niemeyer, C. M. (2010). Mutations of the Shwachman-Bodian-Diamond syndrome gene in patients presenting with refractory cytopenia—do we have to screen?. *haematologica*, 95(4), 689-690.
- Kawakami, T., Mitsui, T., Kanai, M., Shirahata, E., Sendo, D., Kanno, M., ... & Ito, E. (2005). Genetic analysis of Shwachman-Diamond syndrome: phenotypic heterogeneity in patients carrying identical SBDS mutations. *The Tohoku journal of experimental medicine*, 206(3), 253-259.
- Keogh SJ, McKee S, Smithson SF, Grier D, Steward CG. Shwachman-Diamond syndrome: a complex case demonstrating the potential for misdiagnosis as asphyxiating thoracic dystrophy (Jeune syndrome). *BMC Pediatr*. 2012 May 3;12:48.

Kerr, E. N., Ellis, L., Dupuis, A., Rommens, J. M., & Durie, P. R. (2010). The behavioral phenotype of school-age children with shwachman diamond syndrome indicates neurocognitive dysfunction with loss of Shwachman-Bodian-Diamond syndrome gene function. *The Journal of pediatrics*, 156(3), 433-438.

Khan S, Hinks J, Shorto J, Schwarz MJ, Sewell WA. Some cases of common variable immunodeficiency may be due to a mutation in the SBDS gene of Shwachman-Diamond syndrome. *Clin Exp Immunol*. 2008 Mar;151(3):448-54. Epub 2008 Jan 10.

Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus, S., & Ban, N. (2011). Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science*, 334(6058), 941-948.

Kongsuwan, K., Yu, Q., Vincent, A., Frisardi, M. C., Rosbash, M., Lengyel, J. A., & Merriam, J. (1985). A *Drosophila* Minute gene encodes a ribosomal protein. *Nature*, 317(6037), 555.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., ... & Punna, T. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084), 637.

Kuijpers TW, Alders M, Tool AT, Mellink C, Roos D, Hennekam RC. Hematologic abnormalities in Shwachman Diamond syndrome: lack of genotype-phenotype relationship. *Blood*. 2005 Jul 1;106(1):356-61. Epub 2005 Mar 15.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7), 1307-1320.

Lek, Monkol, et al. "Analysis of protein-coding genetic variation in 60,706 humans." *Nature* 536.7616 (2016): 285.

Levinthal, C. (1968). Are there pathways for protein folding?. *Journal de chimie physique*, 65, 44-45.

Liiv, A., & O'Connor, M. (2006). Mutations in the intersubunit bridge regions of 23 S rRNA. *Journal of Biological Chemistry*, 281(40), 29850-29862.

Human insulin market revenue worldwide 2021 | Forecast. (n.d.). Retrieved July 24, 2018, from <https://www.statista.com/statistics/731843/human-insulin-revenue-worldwide/>

Integrated Taxonomic Information System (ITIS) in-line database, Retrieved July 2018, <http://www.itis.gov>

Maaløe, O., & Kjeldgaard, N. O. (1966). Control of macromolecular synthesis: a study of DNA, RNA, and protein synthesis in bacteria.

Mack, D. R., Forstner, G. G., Wilschanski, M. I. C. H. A. E. L., Freedman, M. H., & Durie, P. R. (1996). Shwachman syndrome: exocrine pancreatic dysfunction and variable phenotypic expression. *Gastroenterology*, 111(6), 1593-1602.

Mäkitie O, Ellis L, Durie PR, Morrison JA, Sochett EB, Rommens JM, Cole WG. Skeletal phenotype in patients with Shwachman-Diamond syndrome and mutations in SBDS. *Clin Genet*. 2004 Feb;65(2):101-12. PubMed PMID: 14984468.

Maserati, E., Minelli, A., Pressato, B., Valli, R., Crescenzi, B., Stefanelli, M., ... & Zecca, M. (2006). Shwachman syndrome as mutator phenotype responsible for myeloid dysplasia/neoplasia through karyotype instability and chromosomes 7 and 20 anomalies. *Genes, chromosomes and cancer*, 45(4), 375-382.

Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., ... & Huntemann, M. (2011). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research*, 40(D1), D115-D122.

Marygold, S. J., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Millburn, G. H., ... & Leevers, S. J. (2007). The ribosomal protein genes and Minute loci of *Drosophila melanogaster*. *Genome biology*, 8(10), R216.

Mellink CH, Alders M, van der Lelie H, Hennekam RH, Kuijpers TW. SBDS mutations and isochromosome 7q in a patient with Shwachman-Diamond syndrome: no predisposition to malignant transformation? *Cancer Genet Cytogenet*. 2004 Oct 15;154(2):144-9. 28.

Menne, T. F., Goyenechea, B., Sánchez-Puig, N., Wong, C. C., Tonkin, L. M., Ancliff, P. J., ... & Warren, A. J. (2007). The Shwachman-Bodian-Diamond syndrome protein mediates translational activation of ribosomes in yeast. *Nature genetics*, 39(4), 486.

Munoz, F., Wamser, M., Formont, P., Russel, K., Ross, N., Garmonsway, D., Glur, C. (2018). data.tree: General Purpose Hierarchical Data Structure (Version 0.7.6). Retrieved from <https://CRAN.R-project.org/package=data.tree>

Nakashima E, Mabuchi A, Makita Y, Masuno M, Ohashi H, Nishimura G, Ikegawa S. Novel SBDS mutations caused by gene conversion in Japanese patients with Shwachman-Diamond syndrome. *Hum Genet*. 2004 Mar;114(4):345-8. Epub 2004 Jan 29.

Newman AR, Moghaddam B, Yoon JM. A novel mutation in a Fijian boy with Shwachman-Diamond syndrome. *J Pediatr Hematol Oncol*. 2009 Nov;31(11):847-9.

Nicolis E, Bonizzato A, Assael BM, Cipolli M. Identification of novel mutations in patients with Shwachman-Diamond syndrome. *Hum Mutat*. 2005 Apr;25(4):410.

Nishimura, G., Nakashima, E., Hirose, Y., Cole, T., Cox, P., Cohn, D. H., ... & Unger, S. (2007). The Shwachman–Bodian–Diamond syndrome gene mutations cause a neonatal form of spondylometaphysial dysplasia (SMD) resembling SMD Sedaghatian type. *Journal of medical genetics*, 44(4), e73-e73.

Osterbur, D. L. (2006, May 16). Advanced Course on NCBI Resources: Similarity Searching. Retrieved July 11, 2018, from <https://www.ncbi.nlm.nih.gov/Class/NAWBIS/modules.html>

Parija, S. C., & Jeremiah, S. S. (2013). Blastocystis: Taxonomy, biology and virulence. *Tropical parasitology*, 3(1), 17.

Polacek, N., & Mankin, A. S. (2005). The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Critical reviews in biochemistry and molecular biology*, 40(5), 285-311.

Provost, E., Wehner, K. A., Zhong, X., Ashar, F., Nguyen, E., Green, R., ... & Leach, S. D. (2012). Ribosomal biogenesis genes play an essential and p53-independent role in zebrafish pancreas development. *Development*, 139(17), 3232-3241.

RefSeq growth statistics. (n.d.). Retrieved July 28, 2018, from <https://www.ncbi.nlm.nih.gov/refseq/statistics/>

Rosendahl J, Teich N, Mossner J, Edelmann J, Koch CA. Compound heterozygous mutations of the SBDS gene in a patient with Shwachman-Diamond syndrome, type 1 diabetes mellitus and osteoporosis. *Pancreatology*. 2006;6(6):549-54. Epub 2006 Nov 10.

Savchenko, A., Krogan, N., Cort, J. R., Evdokimova, E., Lew, J. M., Yee, A. A., ... & Kennedy, M. A. (2005). The Shwachman-Bodian-Diamond syndrome protein family is involved in RNA metabolism. *Journal of Biological Chemistry*, 280(19), 19213-19220.

Schaeffer, R. D., & Daggett, V. (2010). Protein folds and protein folding. *Protein Engineering, Design & Selection*, 24(1-2), 11-19.

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, 43(W1), W7-W14.

Shah SS, Bacino CA, Sheehan AM, Shearer WT. Diagnosis of primary immunodeficiency: let your eyes do the talking. *J Allergy Clin Immunol*. 2009 Dec;124(6):1363-4.

Shammas, C., Menne, T. F., Hilcenko, C., Michell, S. R., Goyenechea, B., Boocock, G. R., ... & Warren, A. J. (2005). Structural and mutational analysis of the SBDS protein family Insight into

the leukemia-associated Shwachman-Diamond Syndrome. *Journal of Biological Chemistry*, 280(19), 19221-19229.

Shoji, S., Walker, S. E., & Fredrick, K. (2009). Ribosomal translocation: one step closer to the molecular mechanism. *ACS chemical biology*, 4(2), 93-107.

Shwachman, H., Diamond, L. K., Oski, F. A., & Khaw, K. T. (1964). The syndrome of pancreatic insufficiency and bone marrow dysfunction. *The Journal of pediatrics*, 65(5), 645-663.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., ... & Vauterin, L. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International journal of systematic and evolutionary microbiology*, 52(3), 1043-1047.

Taneichi, H., Kanegane, H., Futatani, T., Otsubo, K., Nomura, K., Sato, Y., ... & Hori, H. (2006). Clinical and genetic analyses of presumed Shwachman-Diamond syndrome in Japan. *International journal of hematology*, 84(1), 60-62.

Tautz, D., Trick, M., & Dover, G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322(6080), 652–656. <https://doi.org/10.1038/322652a0>

Taxonomy Browser. (n.d.). Retrieved August 14, 2018, from <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=2&lvl=3&lin=f&keep=1&srchmode=1&unlock>

Tourlakis, M. E., Zhong, J., Gandhi, R., Zhang, S., Chen, L., Durie, P. R., & Rommens, J. M. (2012). Deficiency of Sbds in the mouse pancreas leads to features of Shwachman–Diamond syndrome, with loss of zymogen granules. *Gastroenterology*, 143(2), 481-492.

Tsangaris, E., Klaassen, R., Fernandez, C. V., Yanofsky, R., Shereck, E., Champagne, J., ... & Abish, S. (2011). Genetic analysis of inherited bone marrow failure syndromes from one prospective, comprehensive and population-based cohort and identification of novel mutations. *Journal of medical genetics*, 48(9), 618-628.

Tymoczko, J. L., Berg, J. M., & Stryer, L. (2011). *Biochemistry: a short course*. Macmillan.

Toiviainen-Salo, S., Mäkitie, O., Mannerkoski, M., Hämäläinen, J., Valanne, L., & Autti, T. (2008). Shwachman–Diamond syndrome is associated with structural brain alterations on MRI. *American Journal of Medical Genetics Part A*, 146(12), 1558-1564.

Toiviainen-Salo S, Raade M, Durie PR, Ip W, Marttinen E, Savilahti E, Mäkitie O. Magnetic resonance imaging findings of the pancreas in patients with Shwachman-Diamond syndrome and mutations in the SBDS gene. *J Pediatr*. 2008 Mar;152(3):434

Torres, A., Cabada, A., & Nieto, J. J. (2003). An exact formula for the number of alignments between two DNA sequences. *DNA Sequence*, 14(6), 427-430.

Valli, R., Pressato, B., Marletta, C., Mare, L., Montalbano, G., Curto, F. L., ... & Maserati, E. (2013). Different loss of material in recurrent chromosome 20 interstitial deletions in Shwachman-Diamond syndrome and in myeloid neoplasms. *Molecular cytogenetics*, 6(1), 56.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), 1189-1191.

Weis, F., Giudice, E., Churcher, M., Jin, L., Hilcenko, C., Wong, C. C., ... & Warren, A. J. (2015). Mechanism of eIF6 release from the nascent 60S ribosomal subunit. *Nature structural & molecular biology*, 22(11), 914.

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API (No. e3179v1). *PeerJ Preprints*.

Woloszynek JR, Rothbaum RJ, Rawls AS, Minx PJ, Wilson RK, Mason PJ, Bessler M, Link DC. Mutations of the SBDS gene are present in most patients with Shwachman-Diamond syndrome. *Blood*. 2004 Dec 1;104(12):3588-90. Epub 2004 Jul 29.

Wong, C. C., Traynor, D., Basse, N., Kay, R. R., & Warren, A. J. (2011). Defective ribosome assembly in Shwachman-Diamond syndrome. *Blood*, blood-2011.

Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology*, 16(1), 18.

Xia J, Bolyard AA, Rodger E, Stein S, Aprikyan AA, Dale DC, Link DC. Prevalence of mutations in ELANE, GFI1, HAX1, SBDS, WAS and G6PC3 in patients with severe congenital neutropenia. *Br J Haematol*. 2009 Nov;147(4):535-42. Epub 2009 Sep 22.

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. D., & Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *science*, 292(5518), 883-896.

Zhang, S., Shi, M., Hui, C. C., & Rommens, J. M. (2006). Loss of the mouse ortholog of the shwachman-diamond syndrome gene (Sbds) results in early embryonic lethality. *Molecular and cellular biology*, 26(17), 6656-6663.

Appendix

Supplementary table 1. List of model organisms used for the PSI-BLAST

Species	Common name
<i>Homo sapiens</i>	Humans
<i>Schizosaccharomyces pombe</i>	Fission yeast
<i>Saccharomyces cerevisiae</i>	Bakers yeast
<i>Caenorhabditis elegans</i>	Roundworm
<i>Mus musculus</i>	Mouse
<i>Rattus Norvegicus</i>	Rat
<i>Dictyostelium discoideum</i>	Slime mold
<i>Sus scrofa</i>	Pig
<i>Arabidopsis thaliana</i>	Thale cress
<i>Archaeoglobus fulgidus</i>	
<i>Xenopus laevis</i>	Frog
<i>Drosophila melanogaster</i>	Fruit fly
<i>Gallus gallus</i>	Chicken
<i>Oryza sativa</i>	Rice
<i>Triticum aestivum</i>	Wheat
<i>Zea mays</i>	Corn
<i>Planaria torva</i>	Planaria