Chapter 3

Results

To fill the gap between knowledge about gene expression and knowledge about the function of those genes, recent approaches focused on investigation of gene expression data in a quantitative manner (see Introduction). A prominent example of those approaches is the comparison of differentially expressed genes in pathological or healthy tissue samples. To this end, those samples need to be as homogeneous as possible, ideally they are micro-dissected. Therefore, information about feedback loops within cells, as well as crosstalk between adjacent cells, tissues and stroma may be lost. Furthermore, the influence of post-translational regulation of the expression level of biologically active proteins, like degradation or protein folding are neglected by those quantitative approaches. The main idea of this thesis is to address these problems by exploiting qualitative instead of quantitative gene expression information on system level, i.e. exploiting expression data of as much different and diverse tissue samples as possible instead of investigating selected samples.

The phenotype angiogenesis is of special therapeutic interest (see Introduction). Its complex modulation includes not only the induction of classical intracellular signaling pathways like hypoxia regulation but also feedback loops within cells, as well as cross-talk between cancer cells, endothelial cells, neighboring tissues and adjacent stroma. Therefore, it represents an ideal candidate for demonstration of the feasibility of the COMMON DENOMINATOR PROCEDURE (CDP). For simplicity, the novel data mining procedure for *in silico* identification of phenotype- or pathway-associated genes is described herein for the phenotype angiogenesis. In the first part, the BASIC COMMON DENOMINATOR PROCEDURE (bCDP) is introduced. In the second part, identification of angiogenesis-specific cDNA libraries is improved by a GENETIC ALGORITHM BASED COMMON DENOMINATOR PROCEDURE gaCDP. Addition-

ally a more sophisticated candidate gene selection is introduced. The third variation, the INDICATOR GENES BASED COMMON DENOMINATOR PROCEDURE IGCDP uses additional biological information in form of well defined indicator genes instead of the genetic algorithm to identify angiogenesis-specific cDNA libraries. Genes are represented by UniGene clusters, whose expression in particular cDNA libraries is listed in the CGAP Expression Data. All COMMON DENOMINATOR PROCEDURES pursue the following scheme:

1. Input data for the procedures are defined.
2. A LIBRARYPROFILE is automatically determined as a set of angiogenesis-specific cDNA libraries;
3. For each human UniGene cluster a GENESCORE is automatically calculated by percentage of libraries from the LIBRARYPROFILE containing it.
4. Candidate genes are automatically selected according to their GENESCORE.
5. Enrichment of angiogenesis-associated genes is determined for internal control and validation of the procedure.

3.1 BASIC CDP

The BCDP is the most elementary variation of the CDP, with a straightforward implementation of the above mentioned scheme. A workflow diagrams for the BCDP is provided in Figure 3.1.

3.1.1 Definition of Input Data

Many sequencing efforts do not routinely extend beyond a limited redundancy in the ESTs, and therefore do not comprehensively reflect the distribution of mRNAs. Therefore, the CGAP Expression Data is rather inhomogeneous. Libraries and sequence sets with reasonable mRNA (EST) numbers, complexity and distribution need to be selected as data source for the procedure. They provide information about the existence of sparsely expressed genes within the tissue sample, which is important for the analysis of co-expression. Therefore, only libraries which met the following criteria were included into the analyses:

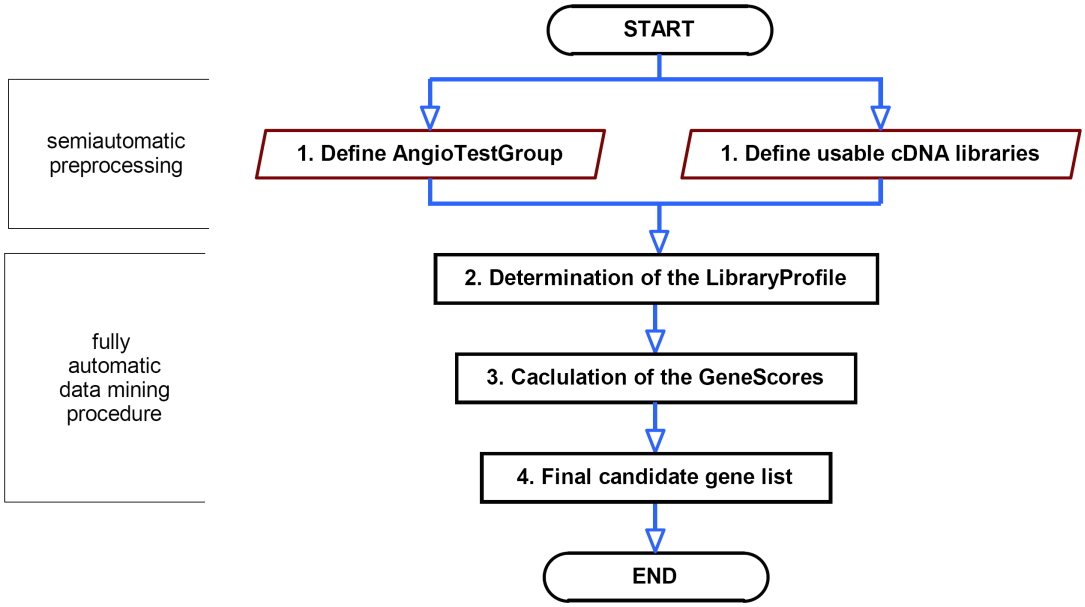


Figure 3.1: Flowchart of the BASIC COMMON DENOMINATOR PROCEDURE. 1. Input data for the procedure are defined including a test set of 170 angiogenesis-associated genes (ANGIOTESTGROUP) as well as a selection of suitable cDNA libraries from CGAP. 2. All libraries are ranked according to the number of UniGene clusters from the ANGIOTESTGROUP contained in them. The top eight libraries containing the highest number of ANGIOTESTGROUP genes compose the set of angiogenesis-specific libraries (LIBRARYPROFILE). 3. A GENESCORE is calculated for all human UniGene clusters, by ranking them according to their frequency of occurrence in the libraries of the LIBRARYPROFILE. 4. All UniGene cluster with a GENESCORE of 100 compose the final candidate gene list.

- The average ratio between the number of ESTs per expressed UniGene cluster observed in that library had to be greater than three.
- Assuming that more than 1000 genes are expressed in a given tissue at a given time [Adams et al., 1995] the libraries had to represent at least 1000 different UniGene clusters.
- Libraries had to represent at most 5000 different UniGene clusters, assuming that larger numbers may be attributed to mixed, pooled, or inhomogeneous samples.

Applying these parameters, 75 libraries out of approximately 7800 cDNA libraries listed in the CGAP Expression Data were identified which matched the inclusion criteria (for more information on the selected cDNA libraries see System and Methods and Table A.1). Restriction to this 75 cDNA libraries concluded the adaptation of the CGAP Expression Data to the COMMON DENOMINATOR PROCEDURES.

The only pathway- or phenotype-specific information necessary for the bCDP is provided with the ANGIOTESTGROUP which consisted of 170 UniGene clusters which are associated with angiogenesis. It was used for selection of an angiogenesis-specific LIBRARYPROFILE and as internal control. To generate the ANGIOTESTGROUP, angiogenesis-associated UniGene clusters were selected semi-automatically based on their GRIF and Gene Ontology annotation using appropriate keywords (see System and Methods).

3.1.2 Determination of the LIBRARYPROFILE

A LIBRARYPROFILE is a small set of angiogenesis-specific cDNA libraries. To obtain those angiogenesis-specific cDNA libraries presence or absence of angiogenesis-associated genes in individual cDNA libraries is determined. In the case of the bCDP the 75 selected cDNA libraries are ranked according to occurrence of different UniGene clusters from the ANGIOTESTGROUP. The eight cDNA libraries containing the highest number of ANGIOTESTGROUP genes composed the LIBRARYPROFILE (see Table 2.1).

3.1.3 Determination of the GENESCORE

For every human UniGene cluster a GENESCORE was calculated based on the LIBRARYPROFILE. This GENESCORE was reflected by the percentage of libraries from the LIBRARYPRO-

FILE containing the particular UniGene cluster. High GENESCORE correlates to expression in most of the selected angiogenesis-specific cDNA libraries, i. e. the higher the GENESCORE, the higher the likeliness of association with angiogenesis. However, UniGene clusters which are present in most or all of the 75 libraries, e. g. housekeeping genes, also achieve *per se* high GENESCORES independent of the LIBRARYPROFILE. This problem was dealt with in the more sophisticated GACDP and IGCDP.

3.1.4 Selection of Candidate Genes

To identify genes that are not yet known to be associated with angiogenesis candidate genes were defined. The 42 UniGene clusters that achieved a GENESCORE of 100 composed the candidate gene list for the BCDP. Extension of this candidate gene list is possible by reducing stringency, i. e. adding UniGene clusters with lower GENESCORES. Table 3.1 shows those genes with a proposed angiogenesis-associated phenotype. These genes can be assigned to various pathways and functionalities that are relevant for angiogenesis. For example, the group 'matrix and motility' includes the known angiogenesis modulator ITGB5 [Nisato et al., 2003]. Other genes which belong to this group (based on their protein properties), but which previously have not been described to be associated with angiogenesis include PLOD and ACTN1. Other functional gene groups include surface receptors and extracellular proteins like BSG [Caudroy et al., 2002] which is also known to modulate angiogenesis. Another surface molecule identified by the procedure is the melanoma antigen CD63 which may be especially interesting due to its cancer-association. Downstream of ligands and receptors are signaling molecules, which also comprise a group of the candidate genes. Prominent among them is the known angiogenesis modulator GRN [Tangkeangsirisin and Serrero, 2004], as well as EWSR1 which is suggested to function as a co-activator for CREB-binding protein (CBP) dependent transcription factors [Araya et al., 2003] (see Table 3.1 for details). The final group is comprised by genes associated with gene expression, i. e. transcription, translation and protein transport. It includes the known angiogenesis modulators WARS [Ewalt and Schimmel, 2002] and HSPB1 [Keezer et al., 2003].

3.1.5 Procedure Control and Validation

The procedure was controlled and validated using three independent methods:

description	name	accession	A	S	P	V
matrix/motility						
procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI)	PLOD	Hs.75093		+		
actinin, alpha 1	ACTN1	Hs.119000				
integrin, beta 5	ITGB5	Hs.149846			+	+
known modulator of angiogenesis [Nisato et al., 2003]						
signaling						
granulin	GRN	Hs.180577			+	+
known modulator of angiogenesis [Tangkeangsirisin and Serrero, 2004]						
Ewing sarcoma breakpoint region 1, transcript variant EWS. *is suggested to function as a co-activator of CREB-binding protein (CBP) dependent transcription factors [Araya et al., 2003], which are known to modulate angiogenesis [Oike et al., 1999].*	EWSR1	Hs.374477				
extracellular/surface						
basigin (OK blood group)	BSG	Hs.501293	+		+	+
known modulator of angiogenesis [Caudroy et al., 2002]						
protein transcription/translation/transport						
heat shock 27kDa protein 1	HSPB1	Hs.76067			+	+
known modulator of angiogenesis [Keezer et al., 2003]						
tryptophanyl-tRNA synthetase	WARS	Hs.82030			+	+
known modulator of angiogenesis [Ewalt and Schimmel, 2002]						
eukaryotic translation elongation factor 1 alpha 1	EEF1A1	Hs.439552		+		
other / unknown						
oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	OGDH	Hs.168669		+		
prosaposin (variant Gaucher disease and variant metachromatic leukodystrophy)	PSAP	Hs.406455		+		
lamin A/C, transcript variant 2.	LMNA	Hs.436441		+		

Table 3.1: Top Candidate Genes of the BASIC CDP with a GENESCORE of 100. A) Genes from the ANGIOTESTGROUP. S) Genes found in the HUVEC proliferation high throughput screen. P) Published angiogenesis modulators. V) Genes linked to vascular biology.

- Enrichment of genes from the ANGIOTESTGROUP (internal procedure control).
- Presence of additional genes known to modulate angiogenesis (procedure validation).
- Presence of experimentally validated, previously unknown angiogenesis-associated genes (procedure validation).

Since the ANGIOTESTGROUP was used for selection of the eight cDNA libraries that composed the LIBRARYPROFILE, enrichment of UniGene clusters from the ANGIOTESTGROUP is not independent of the BCDP and should be observable. While, for that reason, the ANGIOTESTGROUP cannot be used for validation its enrichment within the candidate gene list served as internal control of the procedure. Of the 42 candidate genes one (2.4%) was an ANGIOTESTGROUP gene. Although one hit is at a level below measurement of significance, this hints to an enrichment of genes from the ANGIOTESTGROUP compared to the percentage among all human UniGene clusters represented in the CGAP Expression Data (all UniGene: 0.16%, see System and Methods) or the percentage among all ESTs per UniGene cluster within the restricted data set, i. e. expression data from the 75 selected cDNA

	CDP			CGAP	AngioTestGroup		screen hits	
	basic	GA	IG	#	#	%	#	%
all UniGene	x	x		104,859	170	0.16	611	0.58
			x		73	0.07		
input data	x	x		211,114	1,721	0.82	7129	3.4
			x		1,102	0.52		

Table 3.2: Expectation of AngioTestGroup Genes and Screen Hits. To assess enrichment of AngioTestGroup genes and screen hits their expected number had to be estimated. To this end, their percentage in relation to all human UniGene cluster in the CGAP Expression Data (all UniGene), or in relation to all ESTs per UniGene cluster within the restricted data set (input data) was determined. The number of used AngioTestGroup genes was different in the igCDP (IG) compared to the basic and gaCDP (GA).

libraries, (input data: 0.82 %, see System and Methods and Table 3.2). The different probabilities for getting AngioTestGroup genes by chance is mainly due to the cancer focus of CGAP. As expected, ESTs representing UniGene clusters relevant in angiogenesis were overrepresented in a cancer focused database. Therefore, the latter probability (0.82 %) was better suited to validate the performance of the Common Denominator Procedures, i. e. estimating the accumulation achieved by the Common Denominator Procedures.

To determine significance of enrichment a Yates' corrected χ^2-test [Yates, 1934] was applied using either the probability of getting an AngioTestGroup gene by chance from all UniGene clusters (0.0016) or from the used input data (0.0082). As the probability of drawing a AngioTestGroup gene was quite low in both cases the χ^2-test soon was not applicable due to too small expectation of an AngioTestGroup gene. Generally, it was still observable that significance increased with stringency as did the percentage of genes from the AngioTestGroup. From the 696 UniGene clusters with a GeneScore of at least 75, still 13 (1.9 %) were AngioTestGroup genes. This enrichment was significantly different from the expected number of AngioTestGroup genes even by considering the higher expectation within the used input data ($p<0.005$). This expected behavior can be considered as a positive internal procedure control (see Table 3.3).

In addition to the internal control of the procedure an independent assessment of the candidate gene list was required to validate the procedure and to evaluate potential associations with angiogenesis of the remaining genes including those with so far undefined function. Therefore, the *in silico* defined candidate gene list (without UniGene clusters from the AngioTestGroup) was compared with genes tested positive in a high throughput screen for angiogenesis-factors. This experimental screen had previously been performed at Xantos

s	cand[1]	AngioTestGroup						cand[2]	screen hits					
				all UniGene		input data					all UniGene		input data	
#	#	#	%	fold	χ^2	fold	χ^2	#	#	%	fold	χ^2	fold	χ^2
13	13981	143	1.0	6.3	>99.9	1.2	>99.0	13838	296	2.1	3.6	>99.9	0.6	<exp.
25	7952	103	1.3	8.1	>99.9	1.6	>99.9	7849	231	2.9	5.0	>99.9	0.9	<exp.
38	4871	67	1.4	8.8	>99.9	1.7	>99.9	4804	182	3.8	6.6	>99.9	1.1	>85.0
50	2932	43	1.5	9.4	n.a.	1.8	>99.9	2889	137	4.7	8.1	>99.9	1.4	>99.9
63	1556	25	1.6	10.0	n.a.	2.0	>99.9	1531	96	6.3	10.9	>99.9	1.9	>99.9
75	696	13	1.9	11.9	n.a.	2.3	>99.5	683	57	8.3	14.3	n.a.	2.4	>99.9
88	209	4	1.9	11.9	n.a.	2.3	n.a.	205	23	11.2	19.3	n.a.	3.3	>99.9
100	42	1	2.4	15.0	n.a.	2.9	n.a.	41	5	12.2	21.0	n.a.	3.6	n.a.

Table 3.3: Number of Candidate Genes of the Basic CDP (cand[1]: all candidates; cand[2]: pruned of AngioTestGroup genes) and known angiogenesis-associated genes at different stringency. Stringency is reflected by the minimum GeneScore (s) of a candidate gene. The columns AngioTestGroup and screen hits list the number, estimated enrichment (fold) and significance (according to a χ^2-test, if applicable) of genes present at the respective stringency setting which were members of the AngioTestGroup or of the experimentally defined list of angiogenesis genes. The enrichment was estimated by either considering all human UniGene cluster (all UniGene) in the CGAP Expression Data, or by considering all ESTs per UniGene cluster within our restricted data set (input data).

Biomedicine AG. It was set up with the objective to find secreted factors via a functional genomics approach [Grimm and Kachel, 2002; Koenig-Hoffmann et al., 2005; Zitzler et al., 2004]. For that, proliferation of human umbilical vein endothelial cell (HUVEC) was used to identify candidate genes with functionality in angiogenesis [Denekamp, 1982] (see System and Methods). This XantoScreen™ identified well known angiogenesis-factors such as VEGF or FGF, as well as a list of 466 novel target candidates. To compare these screen hits with the *in silico* candidate gene list a BLAST [Altschul et al., 1990] against all UniGene clusters (identity $\geq 98\,\%$, ≥ 250 nulceotides) was performed. It identified 611 UniGene clusters not member of the AngioTestGroup or the IndicatorGeneSet. The comparison of both lists confirms and extends the specificity of the *in silico* method: additional five (12.2 %) of the remaining 41 candidate genes were members of the experimentally defined list of angiogenesis-associated candidate genes (screen hits). Again, due to too small expectation of a screen hit among the 41 candidate genes a χ^2-test was not applicable. Nevertheless, 12.2 %screen hits compared to expected 0.58 % (all UniGene, see above) or 3.4 % (input data, see above and Table 3.2) indicate a significant enrichment. This fact is strongly supported by the observation, that increased stringency caused a higher percentage of screen hits and as long as applicable a higher significance (smaller p-value). From the 205 UniGene clusters with a GeneScore of at least 88, still 23 (11.2 %) were screen hits. This enrichment was significantly different from the expected number of screen hits even by considering the higher expectation within the used input data ($p<0.001$) (see Table 3.3).

As second means for procedure validation, known angiogenesis modulators that were not part of the AngioTestGroup were identified. To this end, a literature search was performed for the remaining 36 candidate genes. Here, additional four (11.1 %) UniGene clusters were identified that were already known to modulate angiogenesis. Neither the Gene Ontology entries nor the GRIFs of those genes indicated the modulation of angiogenesis, although its association was already known to the public. For that reason, those genes were not part of the AngioTestGroup. With help of Gene Ontology and GRIF it was possible to get a good amount of the known angiogenesis modulators, nevertheless, their annotation is far from complete. Filling these annotation gaps is still a time consuming and highly desirable task. As there is no easy way to identify all known modulators of angiogenesis it was not possible to calculate the significance for identifying these additional five known angiogenesis modulators within the 42 candidate genes.

Altogether nine (22.0 %) of the 41 non-AngioTestGroup candidate genes were either experimentally verified by XantoScreen™ or previously known to modulate angiogenesis.

3.2 Genetic Algorithm Based CDP

Here, the automatic selection of an angiogenesis-specific LibraryProfile was improved by using a genetic algorithm. Additionally, a more sophisticated candidate gene selection was introduced by considering random control profiles and the probabilities of achieving at least the obtained GeneScores by chance, in addition to the GeneScores alone.

3.2.1 Definition of Input Data

The CDP was adapted to CGAP Expression Data, as described above (see bCDP). To this end, 75 cDNA libraries and sequence sets with reasonable mRNA (EST) numbers, complexity and distribution were selected as data source forther procedure from the CGAP Expression Data (see Table *A*.1).

The above described AngioTestGroup was used within the fitness function of the genetic algorithm to select angiogenesis-specific cDNA libraries for the LibraryProfile and again as internal control (see bCDP).

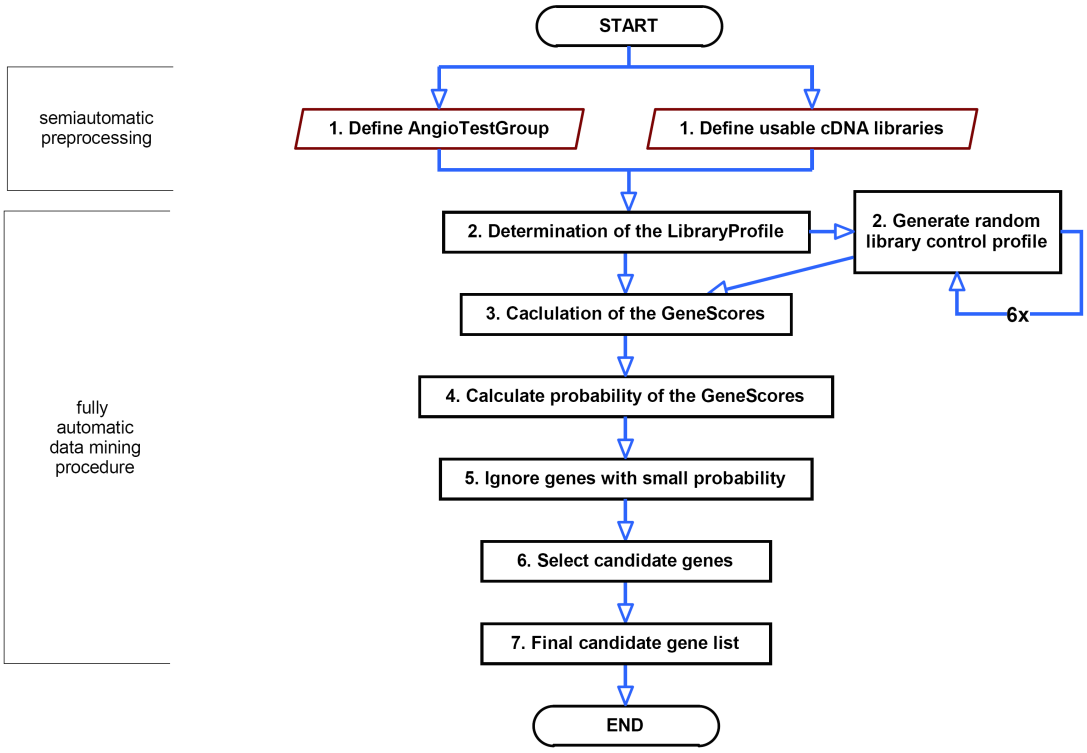


Figure 3.2: Flowchart of the GENETIC ALGORITHM BASED COMMON DENOMINATOR PROCEDURE. 1. Input data for the procedure are defined including a test set of 170 angiogenesis-associated genes (ANGIOTESTGROUP) as well as a selection of suitable cDNA libraries from CGAP. 2. A genetic algorithm is used to determine a set of y angiogenesis-specific libraries (LIBRARYPROFILE). For this LIBRARYPROFILE 6 random library control profiles are generated. To this end, a random library control profile is defined by a set of y randomly selected libraries. 3. For the LIBRARYPROFILE and its corresponding control profiles a GENESCORE is calculated for all human UniGene clusters, by ranking them according to their frequency of occurrence in the libraries of the corresponding profiles. 4. For each UniGene cluster the probability of reaching at least the obtained GENESCORE in the corresponding profile by chance is determined. 5. UniGene clusters with a high probability are ignored in the following steps. 6. Candidate genes are selected based on GENESCORE, probability of the GENESCORE and the random control profiles. 7. The candidate genes are ranked according to their GENESCORE and comprise the final candidate gene list.

3.2.2 Determination of the LibraryProfile

A genetic algorithm was used to select an angiogenesis-specific LibraryProfile containing at least eight cDNA libraries. To this end, a fitness function was designed that maximizes the mean GeneScore of all genes from the AngioTestGroup that were present in the chosen cDNA libraries (see System and Methods for details). Using this approach a set of cDNA libraries was identified which contains as many common AngioTestGroup genes as possible. A UniGene cluster from the AngioTestGroup that occurs only in one of the selected libraries will reduce the mean GeneScore of the AngioTestGroup genes occurring in those libraries. This is in contrast to the bCDP which instead was focused on total number of angiogenesis-associated genes. Comparison of both LibraryProfiles show nevertheless a high similarity. Only one of the again eight cDNA libraries was substituted by the genetic algorithm approach (35627 $\leftrightarrow$ 41605).

3.2.3 Determination of the GeneScore

As described above (see bCDP), for every human UniGene cluster the GeneScore was determined by calculating the percentage of libraries from the LibraryProfile containing it. Generally a higher GeneScore correlates to a higher likeliness of association with angiogenesis. However, UniGene clusters which are present in most or all of the 75 libraries, e. g. housekeeping genes, also achieve *per se* high GeneScores independent of the angiogenesis-specific LibraryProfile. Because these high scoring genes are non-specific and thus not desired, they needed to be identified and removed. Therefore, the probability of achieving at least the obtained GeneScore just by chance was calculated (see System and Methods). To minimize the influence of those UniGene clusters, they were subsequently eliminated from the analyses if they had a probability of more than 36.8 % to have reached their GeneScore by chance (see System and Methods).

3.2.4 Selection of Candidate Genes

In order to identify genes that are not yet known to be associated with angiogenesis candidate genes were defined. To this end, the UniGene clusters were assigned to different ProbabilityGroups, which were defined by common GeneScore and probability of reaching at least that particular GeneScore. ProbabilityGroups with high GeneScore and high

probability of reaching at least that GeneScore were selected based on performance compared to six random library control profiles (see System and Methods). From all UniGene clusters contained in these selected ProbabilityGroups those UniGene clusters were removed, that performed better in at least one of the control profiles. The remaining UniGene clusters compose the final candidate gene list of the gaCDP (see Table 3.4). As a higher GeneScore is still considered to indicate a higher probability of the UniGene cluster to be associated with angiogenesis, the final candidate gene list was ranked by the GeneScore.

Table 3.4 shows a selection from the 22 candidate genes with a GeneScore of 100. As above, these genes were assigned to various pathways and functionalities that are relevant for angiogenesis. For example, the group 'matrix and motility' again includes the known angiogenesis modulator ITGB5 [Nisato et al., 2003]. Other genes which belong to this group (based on their protein properties), but which previously have not been described to be associated with angiogenesis again include PLOD, ACTN1. Other functional gene groups include surface receptors and extracellular proteins like SERPINE1 [Ploplis et al., 2004] which is also known to modulate angiogenesis. Another surface molecule identified by the screen is the melanoma antigen CD63 which may be especially interesting due to its cancer-association. Downstream of ligands and receptors are signaling molecules, which also comprise a group of the candidate genes. Among them are again the known angiogenesis modulator GRN [Tangkeangsirisin and Serrero, 2004], as well as EWSR1 which is suggested to function as a co-activator for CREB-binding protein (CBP) dependent transcription factors [Araya et al., 2003] (see Table 3.1 for details). The final group is comprised by genes associated with gene expression, i. e. transcription, translation and protein transport. It again includes the known angiogenesis modulator WARS [Ewalt and Schimmel, 2002].

3.2.5 Procedure Control and Validation

The procedure was controlled and validated using the same three independent means as described above:

- enrichment of genes from the AngioTestGroup (internal procedure control)
- presence of additional genes known to modulate angiogenesis (procedure validation)

description	name	accession	A	S	P	V
matrix/motility						
integrin beta 5 *known modulator of angiogenesis [Nisato et al., 2003]*	ITGB5	Hs.149846			+	+
actinin alpha 1	ACTN1	Hs.119000				
procollagen-lysine, 2-oxoglutarate 5-dioxygenase	PLOD	Hs.75093		+		
signaling						
Ewing sarcoma breakpoint region 1 *is suggested to function as a co-activator of CREB-binding protein (CBP) dependent transcription factors [Araya et al., 2003], which are known to modulate angiogenesis [Oike et al., 1999].*	EWSR1	Hs.374477				
Granulin *known modulator of angiogenesis [Tangkeangsirisin and Serrero, 2004]*	GRN	Hs.180577			+	+
extracellular/surface						
Serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1 *known modulator of angiogenesis [Ploplis et al., 2004]*	SERPINE1	Hs.414795		+	+	+
protein transcription/translation/transport						
tryptophanyl-tRNA synthetases *known modulator of angiogenesis [Ewalt and Schimmel, 2002]*	WARS	Hs.82030			+	+
others / unknown						
Lamin A/C	LMNA	Hs.436441		+		
Oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	OGDH	Hs.168669		+		

Table 3.4: Top Candidate Genes of the GENETIC ALGORITHM BASED CDP with a GENESCORE of 100. A) Genes from the ANGIOTESTGROUP. S) Genes found in the HUVEC proliferation high throughput screen. P) Published angiogenesis modulators. V) Genes linked to vascular biology.

- presence of experimentally validated, previously unknown angiogenesis-associated genes (procedure validation)

Since the fitness function of the genetic algorithm used the ANGIOTESTGROUP for selection of the LIBRARYPROFILE, the ANGIOTESTGROUP cannot be used for validation. Its enrichment within the candidate gene list again served as internal control of the procedure. 26 (2.8 %) of the 937 candidate genes at the lowest stringency setting, i. e. with a GENESCORE of at least 50, were from the ANGIOTESTGROUP. This correlates to an enrichment of 17.5-fold considering all UniGene clusters or 3.4-fold considering all ESTs from the used input data (for more detail about 'all UniGene' and 'input data' see above, BCDP). This enrichment was significantly different from the expected number of ANGIOTESTGROUP genes even by considering the higher expectation within the used input data ($p<0.001$). This expected behavior can be considered as a positive internal procedure control (see Table 3.5).

In addition to the internal control of the procedure an independent assessment of the candidate gene list was required to validate the procedure and to evaluate potential associations with angiogenesis of the remaining UniGene clusters including ones with so far undefined function. Therefore, the *in silico* defined candidate gene list (without UniGene clusters from the ANGIOTESTGROUP) was compared to genes tested positive in Xantos' high throughput

s	cand[1]	AngioTestGroup						cand[2]	screen hits					
				all UniGene		input data					all UniGene		input data	
#	#	#	%	fold	χ^2	fold	χ^2	#	#	%	fold	χ^2	fold	χ^2
50	937	26	2.8	17.5	n.a.	3.4	>99.9	911	44	4.8	8.3	>99.9	1.4	>97.5
63	578	15	2.6	16.3	n.a.	3.2	n.a.	563	33	5.9	10.2	n.a.	1.7	>99.75
75	291	7	2.4	15.0	n.a.	2.9	n.a.	284	24	8.5	14.7	n.a.	2.5	>99.9
88	94	0	-	-	-	-	-	94	11	11.7	20.2	n.a.	3.4	n.a.
100	22	0	-	-	-	-	-	22	4	18.2	31.4	n.a.	5.4	n.a.

Table 3.5: Number of Candidate Genes of the GENETIC ALGORITHM BASED CDP (cand[1]: all candidates; cand[2]: pruned of ANGIOTESTGROUP genes) and known angiogenesis-associated genes at different stringency. Stringency is reflected by the GENESCORE (s) of a candidate gene. The columns ANGIOTESTGROUP and screen hits list the number, estimated enrichment (fold) and significance (χ^2-test, if applicable) of genes present at the respective stringency setting which were members of the ANGIOTESTGROUP or of the experimentally defined list of angiogenesis genes. The enrichment was estimated by either considering all human UniGene cluster (all UniGene) in the CGAP Expression Data, or by considering all ESTs per UniGene cluster within our restricted data set (input data).

screen for angiogenesis-factors (see above, bCDP). To this end, a BLAST was performed for the clone sequences of the 466 screen hits against all UniGene clusters (identity $\geq 98\,\%$, ≥ 250 nulceotides) identifying 611 UniGene clusters not member of the ANGIOTESTGROUP or the INDICATORGENESET. The comparison of both lists confirms and extends the specificity of the gaCDP at high stringency settings: From the 22 candidate genes with a GENESCORE of 100 four (18.2 %) were members of the experimentally defined list of angiogenesis-associated candidate genes (screen hits). Again, due to too small expectation of a screen hit among the 22 candidate genes a χ^2-test was not applicable. Nevertheless, four screen hits correlate to an enrichment of 5.4-fold considering the used input data and indicate a significant enrichment. This fact is strongly supported by the observation, that increased stringency caused a higher percentage of screen hits and as long as applicable a higher significance (smaller p-value). From the 197 UniGene clusters with a GENESCORE of at least 75, still 13 (6.6 %) were screen hits. This enrichment was significantly different from the expected number of screen hits even by considering the higher expectation within the used input data ($p<0.025$) (see Table 3.5). At the lowest stringency settings, i. e. 50 and 63, a significant enrichment of screen hits is not observable.

As second means for procedure validation, known angiogenesis modulators that were not part of the ANGIOTESTGROUP were identified. To this end, a literature search was performed for the remaining 18 candidate genes (without screen hits) at the highest stringency setting, i. e. GENESCORE of 100. Here additional three (16.6 %) UniGene clusters were identified that were already known to modulate angiogenesis. Again, neither the Gene Ontology entries nor the GRIFs of those genes indicated the modulation of angiogenesis, although its association

was already known to the public. As there is no easy way to identify all known modulators of angiogenesis it was not possible to calculate the significance for identifying additional three known angiogenesis modulators within the 18 candidate genes.

Altogether seven (31.8 %) of the 22 candidate genes with a GeneScore of 100 were either experimentally verified by XantoScreen™ or previously known to modulate angiogenesis.

3.3 Indicator Genes Based CDP

The automatic selection of an angiogenesis-specific LibraryProfile was improved by adding more biological information in form of a small set of manually selected indicator genes (IndicatorGeneSet). In contrast to the above described Common Denominator Procedures, a multitude of LibraryProfiles was generated, each biased by using different combinations of those indicator genes. The semiautomatically selected AngioTestGroup was used to identify the most angiogenesis-specific LibraryProfiles as well as for determination of candidate genes as introduced for the gaCDP. Scalability of the CDP was improved by ranking these candidate genes according to their multiplicity of occurrence in those angiogenesis-specific profiles.

3.3.1 Definition of Input Data

The CDP was adapted to CGAP Expression Data, as described above (see bCDP). To this end, 75 cDNA libraries and sequence sets with reasonable mRNA (EST) numbers, complexity and distribution were selected as data source for the procedure from the CGAP Expression Data (see Table *A*.1).

A subset of the above described AngioTestGroup was used for the selection of angiogenesis-specific LibraryProfiles and as internal control. To increase specificity and to reduce noise effects the 170 angiogenesis modulators were restricted to 73 UniGene clusters which occur in at least eight and at most 37 (50 %) of the 75 used cDNA libraries (see System and Methods).

Additionally, a set of well defined indicator genes were selected, that modulate angiogenesis. This IndicatorGeneSet consisted of the following six known pro-angiogenic factors [Shoshani et al., 2002; Stoeltzing et al., 2003; Tsukagoshi et al., 2003; Wiesener et al., 1998]:

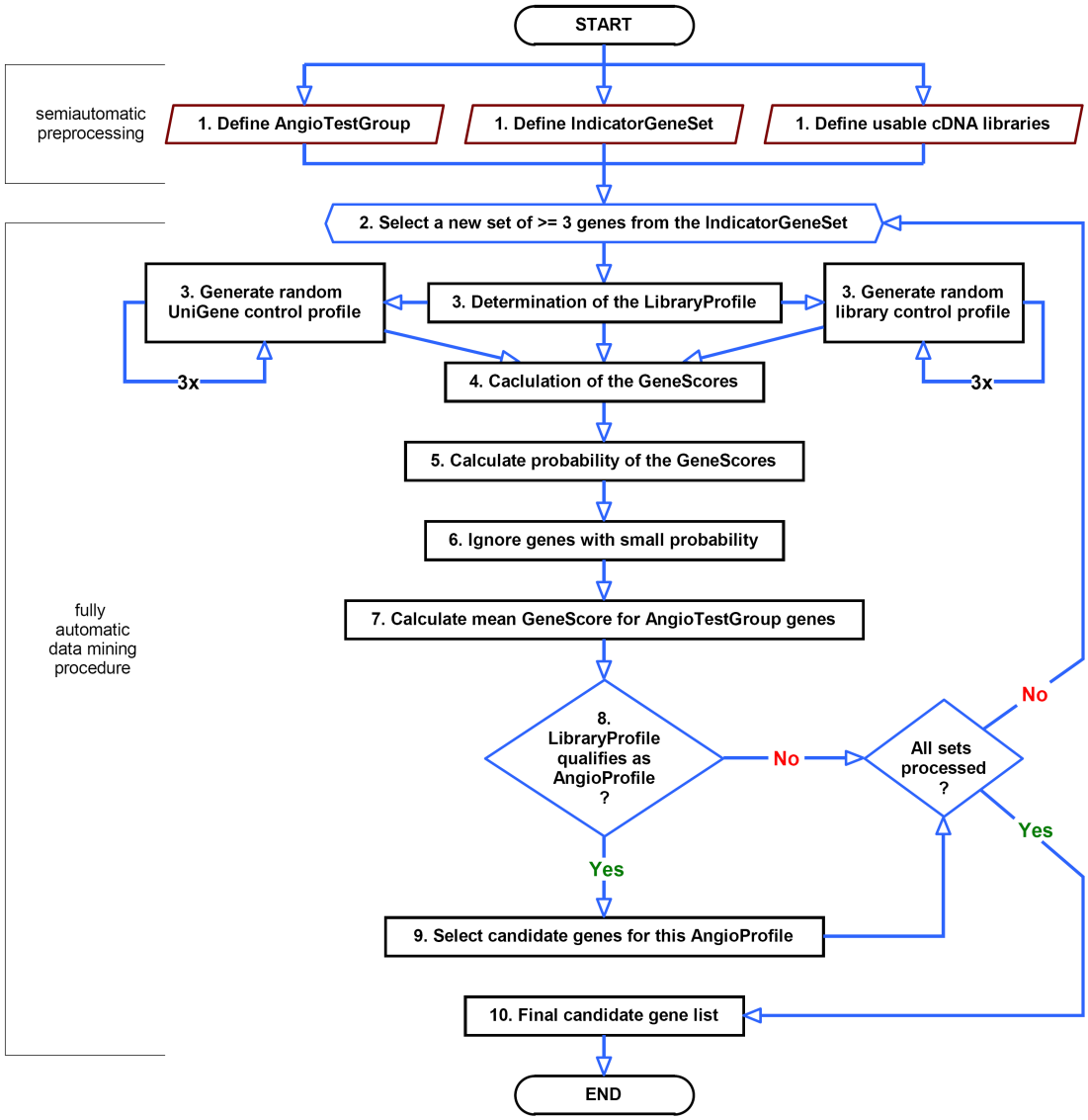


Figure 3.3: Flowchart of the INDICATOR GENES BASED COMMON DENOMINATOR PROCEDURE. 1. Input data for the procedure are defined.2. From all combinations of ≥ 3 indicator genes, a not yet processed set of x genes is selected. 3. This set is used to determine a set of y libraries (LIBRARYPROFILE) with common or similar presence/absence pattern. For this LIBRARYPROFILE 3 random UniGene control profiles and 3 random library control profiles are generated. 4. For the LIBRARYPROFILE and its corresponding control profiles a GENESCORE is calculated for all human UniGene clusters, by ranking them according to their frequency of occurrence in the libraries of the corresponding profiles. 5. For each UniGene cluster the probability of reaching at least the obtained GENESCORE in the corresponding profile by chance is determined. 6. UniGene clusters with a high probability are ignored in the following steps. 7. The mean GENESCORE of UniGene clusters from the ANGIOTESTGROUP is calculated for the LibraryProfile and its corresponding control profiles. 8. The LIBRARYPROFILE qualifies as ANGIOPROFILE if its mean value is greater than the mean value of all corresponding control profiles. 9. Candidate genes are selected based on GENESCORE, probability of the GENESCORE and the random control profiles. Steps 2-8 are repeated for all combinations of ≥ 3 genes from the INDICATORGENESET. 10. All candidate genes are ranked according to their multiplicity of occurrence in the ANGIOPROFILES and comprise the final candidate gene list.

- Hypoxia-inducible factor 1 alpha (HIF1A)
- Hypoxia-inducible factor 1-responsive gene (DDIT4)
- Vascular endothelial growth factor (VEGF)
- Insulin-like growth factor 1 receptor (IGFR1)
- Endothelial cell growth factor 1 (ECGF1)
- Endothelial PAS domain protein 1 (EPAS1)

3.3.2 Determination of the LibraryProfile

A LibraryProfile is a small set of angiogenesis-specific cDNA libraries. To obtain those angiogenesis-specific cDNA libraries presence or absence of genes from the IndicatorGeneSet in individual cDNA libraries is determined. All possible combinations of three to six of these indicator genes were used to generate 42 individual LibraryProfiles (see System and Methods). Those LibraryProfiles represent a set of libraries with common occurrence of the chosen indicator genes.

3.3.3 Determination of the GeneScore

As described above (see bCDP), for every LibraryProfile and every human UniGene cluster the GeneScore was determined by calculating the percentage of libraries from the LibraryProfile containing that particular UniGene cluster. Generally a higher GeneScore correlates to a higher likeliness of association with angiogenesis. However, UniGene clusters which are present in most or all of the 75 used cDNA libraries, e. g. housekeeping genes, also achieve *per se* high GeneScores independent of the angiogenesis-specific LibraryProfile. Because these high scoring genes are non-specific and thus not desired, they needed to be identified and removed. Therefore, the probability of achieving at least the obtained GeneScore just by chance was calculated (s_{imp}, see System and Methods). To minimize the influence of those UniGene clusters, all clusters which had a probability of more than 36.8 % ($s_{imp} \leq 1.0$) to have reached their GeneScore by chance were subsequently eliminated from the analyses (see System and Methods).

3.3.4 Selection of AngioProfiles

In this step the most suitable LibraryProfiles for the identification of angiogenesis-associated genes were selected from all 42 LibraryProfiles. The rational underlying the selection is that genes from the AngioTestGroup should obtain higher GeneScores in the desired LibraryProfiles compared to random control profiles. This assessment of LibraryProfiles by application of an independent (more objective) dataset related to angiogenesis is of particular importance as the procedure starts with a user-defined (i. e. subjective) IndicatorGeneSet. Co-occurrence of the chosen pro-angiogenic indicator genes in those selected libraries hints toward the angiogenesis-specificity of the LibraryProfile. Nevertheless, not all combinations of the indicator genes may jointly participate in pathways that lead to the phenotype angiogenesis. Additionally, even if they do, this fact may not be represented within the CGAP Expression Data. Therefore, the most angiogenesis-specific LibraryProfiles were selected for further processing.

For every LibraryProfile three random UniGene cluster control profiles and three random library control profiles were created (see System and Methods). Then, the mean value of the GeneScore over all AngioTestGroup genes was calculated for each LibraryProfile and its corresponding six control profiles (see System and Methods, and Table 3.6). In summary, the average of these mean values was already greater over all 42 LibraryProfiles (49.6 %) compared to the 252 control profiles (random library: 38.7 %, random UniGene: 41.1 %, combined: 39.9 %). Of all LibraryProfiles 16 had a higher mean value than any of their corresponding control profiles. These were considered to be most angiogenesis-specific and thus termed AngioProfiles. As expected, the removal of unspecific LibraryProfiles lead to a further increased average mean value over all 16 AngioProfiles (52.9 %).

3.3.5 Selection of Candidate Genes

To identify genes that are not yet known to be associated with angiogenesis candidate genes were defined for each of the 16 AngioProfile as described for the gaCDP. To this end, the GeneScore, its probability and control profiles were considered (see System and Methods). Multiple occurrence of the same candidate gene in different AngioProfiles indicates a higher probability of the corresponding UniGene cluster to be associated with angiogenesis. Therefore, candidate genes were ranked according to their multiplicity of occurrence in the

IndicatorGeneSet						CGAP	AngioTestGroup genes with $s_{imp} \geq 1$		
1	2	3	4	5	6	# libraries	# profile	# controls	δ mean GeneScore
x	x	x			x	26	3	6	+9.0
x		x	x		x	9	16	11	+6.3
	x	x		x	x	19	10	2	+6.1
x		x	x	x	x	16	17	1	+5.9
		x	x	x	x	21	8	2	+4.4
x		x		x	x	9	27	10	+3.1
x		x	x	x		28	3	6	+3.1
x	x			x	x	26	4	1	+3.0
x			x		x	21	7	15	+2.0
x	x	x	x	x	x	21	8	5	+2.0
		x		x	x	15	24	6	+1.3
x		x			x	21	7	1	+1.1
		x	x	x		9	20	2	+1.1
	x	x	x	x	x	26	3	5	+1.1
	x	x	x		x	19	11	10	+1.0
	x		x		x	11	17	20	+0.6

Table 3.6: Detailed Information for the Selected AngioProfiles. It shows the UniGene clusters from the IndicatorGeneSet that were used for generation of the AngioProfiles (1 = HIF1 responsive, 2 = IGFR1, 3 = HIF1A, 4 = ECGF1, 5 = VEGF, 6 = EPAS1) and the number of cDNA libraries composing these AngioProfiles (# libraries). Also shown is the number of AngioTestGroup genes with $s_{imp} \geq 1$ within the AngioProfiles (# profile) and its corresponding six control profiles (# controls) as well as the difference of the mean GeneScore of those AngioTestGroup genes from the AngioProfile compared to the highest mean GeneScore of the corresponding six control profiles (δ mean GeneScore).

AngioProfiles yielding the final candidate gene list.

Table 3.7 shows a list of genes that were identified by the procedure as genes with a proposed angiogenesis-associated phenotype. These genes can be assigned to various pathways and functionalities that are relevant for angiogenesis. For example, the group 'matrix and motility' includes known angiogenesis related genes ITGB5 [Nisato et al., 2003], FN1 [Krishnamachary et al., 2003] CAPN1 [Su et al., 2004], LAMA5 [Sasaki and Timpl, 2001] and ADAM15 [Horiuchi et al., 2003; Trochon-Joseph et al., 2004]. Interestingly (or rather consequently) the *in silico* screen also points to LGALS3BP as being angiogenesis-associated. LGALS3BP is a protein which binds FN1, ITGB5 and COL6A2 [Marchetti et al., 2002]. COL6A2 is also known to modulate angiogenesis [Daniels et al., 1996; Iyengar et al., 2003] and was a candidate gene in four of the 16 AngioProfiles. Thus LGALS3BP fits well into this group. Other genes which belong to this group (based on their protein properties), but which previously have not been described to be associated with angiogenesis include TUBB-5, ACTN1 and PLOD. Other functional gene groups include soluble factors (ligands: TGFBI [Thorey et al., 2004] and SVAP1 [Davis et al., 2002] both associated with vascular biology), and surface receptors (CD151 [Wright et al., 2004]) which are known to

modulate angiogenesis. Downstream of ligands and receptors are signaling molecules and molecules related to gene expression, which also comprise a group of the candidate genes (see Table 3.7 for details). Of interest for further research are also those candidate genes for which so far no function has been assigned. Among those 'unknowns' the candidate genes OS-9 and PML (see Table 3.7) may be particularly exciting, because these have been described as cancer-associated genes. Thus, these genes may contribute to or modulate tumor angiogenesis.

3.3.6 Procedure Control and Validation

The procedure was controlled and validated using the same three independent means as described above:

- Enrichment of genes from the AngioTestGroup and IndicatorGeneSet (internal procedure control).
- Presence of additional genes known to modulate angiogenesis (procedure validation).
- Presence of experimentally validated, previously unknown angiogenesis-associated genes (procedure validation).

Since the IndicatorGeneSet was used for selection of the 42 LibraryProfile and the AngioTestGroup for selection of the 16 AngioProfiles, the AngioTestGroup and the IndicatorGeneSet were not independent of the procedure and cannot be used for validation. While they cannot be used for validation their enrichment within the candidate gene list served as internal control of the procedure. The above described candidate gene list harbored 2031 candidate genes if one considered the 'lowest' specificity setting, i. e. definition as candidate in at least one of the 16 AngioProfiles. Five of the six UniGene clusters from the IndicatorGeneSet were contained in those 2031 candidate genes. The individual indicator genes were identified as candidate gene by up to ten different AngioProfiles (IGFR1: 0, EGF1: 3, VEGF: 4, HIF1A: 7, EPAS1: 10 and DDIT4: 10). At this specificity setting 18 (0.89 %) UniGene clusters from the AngioTestGroup were contained in the 2031 candidate genes. This correlates to an enrichment of 12.7-fold AngioTestGroup genes considering all UniGene clusters (0.07 %) or 1.7-fold considering all ESTs from the used input data (0.52 %, see above and Methods). The specificity for

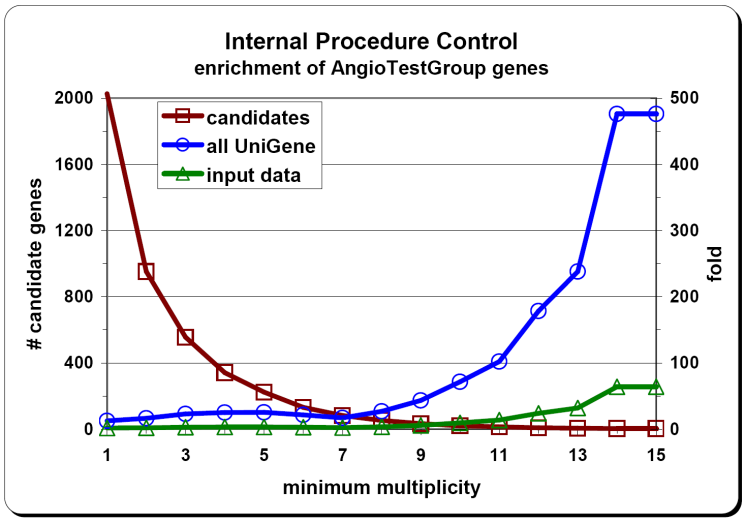
description	name	accession	I	A	S	P	V
matrix/motility							
fibronectin 1 *known modulator of angiogenesis [Krishnamachary et al., 2003]*	FN1	Hs.418138			+	+	+
integrin beta 5 *known modulator of angiogenesis [Nisato et al., 2003]*	ITGB5	Hs.149846				+	+
lectin galactoside-binding soluble 3 binding protein (LGALS3BP, MAC2-BP) *binds COL6A2, FN1 and ITGB5 [Marchetti et al., 2002].*	LGALS3BP	Hs.79339					
calpain 1 *known modulator of angiogenesis [Su et al., 2004]*	CAPN1	Hs.356181				+	+
laminin alpha 5 *known modulator of angiogenesis [Sasaki and Timpl, 2001]*	LAMA5	Hs.11669				+	+
zyxin *is modulated by an orphan type I receptor of transforming growth factor beta (ALK-1), which is believed to be implicated in angiogenesis [Lamouille et al., 2002].*	ZYX	Hs.75873					
tubulin beta 5	TUBB-5	Hs.274398					
actinin alpha 1	ACTN1	Hs.119000					
procollagen-lysine, 2-oxoglutarate 5-dioxygenase	PLOD	Hs.75093			+		
a disintegrin and metalloproteinase domain 15 (metargidin) *known modulator of angiogenesis [Horiuchi et al., 2003; Trochon-Joseph et al., 2004]*	ADAM15	Hs.312098			+	+	+
signaling							
endothelial PAS domain protein 1 *known modulator of angiogenesis [Wiesener et al., 1998]*	EPAS1	Hs.8136	+	+		+	+
HIF-1 responsive gene *known modulator of angiogenesis [Shoshani et al., 2002]*	DDIT4	Hs.111244	+	+	+	+	+
cyclin-dependent kinase inhibitor 1A (CDKN1A, CIP1, p21) *known modulator of angiogenesis [Koshiji and Huang, 2004; Kumar and Vadlamudi, 2002]*	CDKN1A	Hs.370771				+	+
mitogen-activated protein kinase-3 11(MAP3K11, MLK3) *suppresses the activation of the angiogenesis modulator MAPK14 [Chadee and Kyriakis, 2004; Xu et al., 2004].*	MAP3K11	Hs.432787					
Ewing sarcoma breakpoint region 1 *is suggested to function as a co-activator of CREB-binding protein (CBP) dependent transcription factors [Araya et al., 2003].*	EWSR1	Hs.374477					
cAMP responsive element binding protein 3 *CREB-binding proteins are known to modulate angiogenesis [Oike et al., 1999]*	CREB3	Hs.287921			+		
soluble factors							
soluble vascular adhesion protein 1 *may be linked to angiogenesis [Davis et al., 2002]*	SVAP1	Hs.325081					+
beta-induced transforming growth factor *known modulator of angiogenesis [Thorey et al., 2004]*	TGFBI	Hs.421496		+		+	+
extracellular/surface							
CD151 antigen *known modulator of angiogenesis [Wright et al., 2004]*	CD151	Hs.512857				+	+
protein transcription/translation/transport							
tryptophanyl-tRNA synthetases *known modulator of angiogenesis [Ewalt and Schimmel, 2002]*	WARS	Hs.82030				+	+
LAG1 longevity assurance homolog 2 (S. cerevisiae), transcript variant 3	LASS2	Hs.285976			+		
MCM6 minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae) *plays essential role during proliferation of vascular smooth muscle cells [Bruemmer et al., 2003]*	MCM6	Hs.444118					+
other / unknown							
Lamin A/C	LMNA	Hs.436441			+		
amplified in osteosarcoma	OS-9	Hs.76228					
promyelocytic leukemia, transcript variant 8	PML	Hs.89633					
hypothetical protein 628	LOC56270	Hs.201390			+		

Table 3.7: Top Candidate Genes of the INDICATOR GENES BASED CDP occurring in at least eight profiles. I) Genes from the INDICATORGENESET; A) Genes from the ANGIOTESTGROUP. S) Genes found in the HUVEC proliferation high throughput screen. P) Published angiogenesis modulators. V) Genes linked to vascular biology.

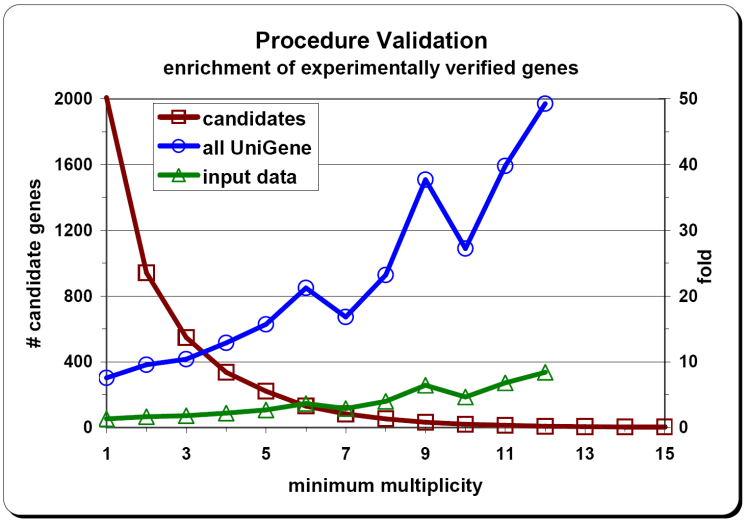
m	cand[1]	AngioTestGroup						cand[2]	screen hits					
				all UniGene		input data					all UniGene		input data	
#	#	#	%	fold	χ^2	fold	χ^2	#	#	%	fold	χ^2	fold	χ^2
1	2026	18	0.9	12.7	n.a.	1.7	>95.0	2008	88	4.4	7.6	>99.9	1.3	>98.0
2	952	11	1.2	16.6	n.a.	2.2	n.a.	941	52	5.5	9.5	>99.9	1.6	>99.9
3	556	9	1.6	23.1	n.a.	3.1	n.a.	547	33	6.0	10.4	n.a.	1.8	>99.75
4	341	6	1.8	25.1	n.a.	3.4	n.a.	335	25	7.5	12.9	n.a.	2.2	>99.9
5	224	4	1.8	25.6	n.a.	3.4	n.a.	220	20	9.1	15.7	n.a.	2.7	>99.9
6	132	2	1.5	21.7	n.a.	2.9	n.a.	130	16	12.3	21.2	n.a.	3.6	n.a.
7	83	1	1.2	17.1	n.a.	2.3	n.a.	82	8	9.8	16.8	n.a.	2.9	n.a.
8	53	1	1.9	27.0	n.a.	3.6	n.a.	52	7	13.5	23.2	n.a.	4.0	n.a.
9	33	1	3.0	43.3	n.a.	5.8	n.a.	32	7	21.9	37.7	n.a.	6.4	n.a.
10	20	1	5.0	71.4	n.a.	9.6	n.a.	19	3	15.8	27.2	n.a.	4.6	n.a.
11	14	1	7.1	102.0	n.a.	13.7	n.a.	13	3	23.1	39.8	n.a.	6.8	n.a.
12	8	1	12.5	178.6	n.a.	24.0	n.a.	7	2	28.6	49.3	n.a.	8.4	n.a.
13	6	1	16.7	238.1	n.a.	32.1	n.a.	5	0	-	-	n.a.	-	n.a.
15	3	1	33.3	476.1	n.a.	64.1	n.a.	2	0	-	-	n.a.	-	n.a.

Table 3.8: Number of Candidate Genes of the INDICATOR GENES BASED CDP (cand[1]: pruned of indicator genes; cand[2]: pruned of indicator genes and ANGIOTESTGROUP genes) and known angiogenesis-associated genes at different stringency. Stringency is reflected by the minimum multiplicity (m) of a candidate gene in ANGIOPROFILES, i.e. occurring in at least that number of ANGIOPROFILES. The columns ANGIOTESTGROUP and screen hits list the number, estimated enrichment (fold) and significance (according to a χ^2-test, if applicable) of genes present at the respective stringency setting which were members of the ANGIOTESTGROUP or of the experimentally defined list of angiogenesis genes. The enrichment was estimated by either considering all human UniGene cluster (all UniGene) in the CGAP Expression Data, or by considering all ESTs per UniGene cluster within our restricted data set (input data).

angiogenesis increased further upon ranking of candidate genes according to multiplicity of occurrence in distinct ANGIOPROFILES: For example from 55 candidate genes that were present in eight or more ANGIOPROFILES, still two (3.6 %) were indicator genes and one (1.8 %) was an ANGIOTESTGROUP genes. The latter percentage correlates to an enrichment of 27-fold considering all UniGene and 3.6-fold considering the used input data. A graphical representation of the enrichment at different stringency setting is shown in Figure 3.4. Due to very small probability of getting a ANGIOTESTGROUP gene by chance (0.0007 or 0.0052), a χ^2-test was not applicable for high specificity settings. Nevertheless, for the 18 ANGIOTESTGROUP genes of the 2031 candidate genes that occurred in at least one ANGIOPROFILES statistical significance could be shown ($p<0.005$) for the input data related estimation. In summary, the expected and observed enrichment of UniGene clusters from the INDICATORGENESET and ANGIOTESTGROUP can be considered as positive internal control. From the above mentioned 55 candidate genes additional 14 (25.5 %) were not part of the ANGIOTESTGROUP or the INDICATORGENESET, but were known to be related to angiogenesis, as concluded from extended literature analysis and from wet-lab experiments as described below (see Table 3.8).



(a) internal procedure control



(b) procedure validation

Figure 3.4: Validation of the INDICATOR GENES BASED COMMON DENOMINATOR PROCEDURE. The number (left y-axis, open square) and enrichment (right y-axis) of candidate genes (3.4(a): ANGIOTESTGROUP 3.4(b): experimental screening hits from Xantos' angiogenesis screen) is shown in relation to the stringency setting of the procedure. Stringency (x-axis) is defined by the minimum multiplicity of a candidate gene in ANGIOPROFILES, i. e. occurring in at least that number of ANGIOPROFILES. The enrichment was estimated by either considering all human UniGene cluster in the CGAP Expression Data (open circle), or by considering all ESTs per UniGene cluster within the restricted data set (open triangle). Note that with increasing stringency setting the number of candidate genes decreases and the enrichment of angiogenesis-associated genes increases. This is expected for a procedure that specifically enriches angiogenesis genes.

In addition to the internal control of the procedure an independent assessment of the candidate gene list was required to validate the procedure and to evaluate potential associations with angiogenesis of the remaining genes including ones with so far undefined function. Therefore, the *in silico* defined candidate gene list (without UniGene clusters from the IndicatorGeneSet and the AngioTestGroup) was compared to genes tested positive in Xantos' high throughput screen for angiogenesis-factors (see above, bCDP). To this end, a BLAST was performed for the 466 *in vivo* candidate genes against all UniGene clusters (identity $\geq 98\,\%$, ≥ 250 nulceotides) identifying 611 UniGene clusters not member of the AngioTestGroup or the IndicatorGeneSet. The comparison of both lists confirms and extends the specificity of the igCDP: applying the 'lowest' specificity setting, i. e. definition as candidate in at least one of the 16 AngioProfiles, 88 (4.4 %) of all candidate genes were members of the experimentally defined list of angiogenesis-associated candidate genes (screen hits). This fraction of genes that were also experimentally found to be associated with angiogenesis increased to seven UniGene clusters (13.5 %) for the higher specificity setting (candidates present in eight or more AngioProfiles). Again, due to too small expectation of a screen hit among the 52 candidate genes at the higher specificity setting a χ^2-test was not applicable. Nevertheless, seven screen hits correlate to an enrichment of 4.0-fold considering the used input data and indicate a significant enrichment. This fact is strongly supported by the observation, that increased stringency caused a higher percentage of screen hits and as long as applicable a higher significance (smaller p-value). From the 220 candidate genes that occurred in at least five AngioProfiles still 20 (9.1 %) were screen hits. This enrichment was significantly different from the expected number of screen hits even by considering the higher expectation within the used input data ($p<0.001$) (see Table 3.8). Even at the lowest stringency settings, (occurrence in at least one AngioProfile) the enrichment of screen hits was significant ($p<0.002$).

As second means for procedure validation, known angiogenesis modulators that were not part of the IndicatorGeneSet or the AngioTestGroup were identified. To this end, a literature search was performed for the remaining 45 candidate genes that occurred in at least eight AngioProfiles (without screen hits and UniGene clusters from the IndicatorGeneSet and the AngioTestGroup). Here additional six (13.3 %) UniGene clusters were identified that were already known to modulate angiogenesis. Again, neither the Gene Ontology entries nor the GRIFs of those genes indicated the modulation of angiogenesis, al-

m	candidates	IndicatorGeneSet		AngioTestGroup		screen hits		published		total	
#	#	#	%	#	%	#	%	#	%		
1	2031	5	0.2	18	0.9	88	4.4	n.d.	-	111	5.5
2	957	5	0.5	11	1.2	52	5.5	n.d.	-	68	7.1
3	561	5	0.9	9	1.6	33	5.9	n.d.	-	47	8.4
4	345	4	1.2	6	1.7	25	7.3	n.d.	-	35	10.1
5	227	3	1.3	4	1.8	20	8.9	n.d.	-	27	11.9
6	135	3	2.2	2	1.5	16	11.9	n.d.	-	21	15.6
7	86	3	3.5	1	1.2	8	9.4	n.d.	-	12	14.0
8	55	2	3.6	1	1.8	7	13.0	6	10.9	16	29.1
9	35	2	5.7	1	2.9	7	20.6	4	11.4	14	40.0
10	22	2	9.1	1	4.6	3	14.3	2	9.1	8	36.4
11	14	0	-	1	7.1	3	21.4	2	14.3	6	42.9
12	8	0	-	1	12.5	2	25.0	2	25.0	5	62.5
13	6	0	-	1	16.7	0	-	2	33.3	3	50.0
15	3	0	-	1	33.3	0	-	0	-	1	33.3

Table 3.9: Known Modulators of Angiogenesis for the Indicator Genes Based CDP at different stringency. Stringency is reflected by the minimum multiplicity (m) of a candidate gene in AngioProfiles, i.e. occurring in at least that number of AngioProfiles. The columns IndicatorGeneSet, AngioTestGroup, screen hits and published list the number of genes present at the respective stringency setting which were members of the AngioTestGroup, of the experimentally defined list of angiogenesis genes or which were published modulators of angiogenesis, but not member of the previous groups. For candidate genes that occurred in less than eight AngioProfile no literature search was performed. The last column (total) lists the sum of all before mentioned validated angiogenesis modulators.

though its association was already known to the public. As there is no easy way to identify all known modulators of angiogenesis it was not possible to calculate the significance for identifying additional six known angiogenesis modulators within the 45 remaining candidate genes.

At this stringency setting, altogether 13 (25.0 %) of the 52 non-AngioTestGroup, non-IndicatorGeneSet candidate genes that occurred in at least eight AngioProfiles were either experimentally verified by XantoScreen™ or previously known to modulate angiogenesis. This enrichment increased further to 57.1 % for the seven non-AngioTestGroup, non-IndicatorGeneSet candidate genes that occurred in at least twelve AngioProfiles (see Table 3.9).

3.4 Summary

For the validation of the procedure and estimation of the specificity of the candidate gene lists, three independent parameters were determined:

1. **Internal control:** enrichment of genes from the AngioTestGroup and the Indi-

CATORGENESET.

2. **Experimental validation:** presence of experimentally validated angiogenesis-associated genes.

3. **Literature validation:** presence of additional genes known to modulate angiogenesis.

3.4.1 Internal Procedure Control

UniGene clusters from the ANGIOTESTGROUP, and in case of the IGCDP also UniGene clusters from the INDICATORGENESET, were used for selection of angiogenesis-specific LIBRARYPROFILES and ANGIOPROFILES. Therefore, enrichment of those UniGene clusters within the candidate gene list would be expected. All three procedures identified a significant amount of those UniGene clusters within their final candidate gene lists, compared to the relatively small probability of their occurrence. This probability was estimated using two different approaches (see Table 3.2).

- All UniGene: the probability of drawing an ANGIOTESTGROUP gene from all human UniGene clusters present in the CGAP Expression Data.

- Input data: the probability of drawing an EST corresponding to an ANGIOTESTGROUP gene from all ESTs within the restricted data set (75 libraries).

CGAP Expression Data has a cancer focus and is thus biased towards angiogenesis. This is a possible explanation for the higher probability within the restricted data set (input data). Therefore, it forms a more conservative estimation of the expected number of ANGIOTESTGROUP genes. For all three COMMON DENOMINATOR PROCEDURES a significant enrichment of UniGene clusters from the ANGIOTESTGROUP could be shown even at the lowest stringency settings (small minimum GENESCORE or small multiplicity, i. e. occurrence in multiple ANGIOPROFILES). At high stringency settings the χ^2-test was not applicable, due to the low expectation of ANGIOTESTGROUP genes within the small number of candidate genes. Nevertheless, the percentage, and thus the estimated enrichment, of UniGene clusters from the ANGIOTESTGROUP increased with higher stringency settings. The BCDP enriched ANGIOTESTGROUP genes by up to three-fold (input data), the GACDP by up to three-fold (input data) and the IGCDP from two-fold to 64-fold (input data). As these

enrichments could not all be statistically confirmed (see above) their exact value must be treated with caution. Nevertheless, they indicate an enrichment and can be considered as a positive internal control.

3.4.2 Procedure Validation - Experimental

Presence of experimentally validated angiogenesis-associated genes provided further independent means for validation of the candidate gene lists. Xantos previously performed a high throughput screen for the identification of angiogenesis, based upon a HUVEC proliferation assay. Some of these screen hits were previously known to modulate angiogenesis, for many association with angiogenesis was yet not described. The candidate genes were compared to the 611 UniGene clusters that were mapped to the 466 hits of the proliferation screen (without Unigene clusters from the AngioTestGroup or the IndicatorGeneSet). For all three Common Denominator Procedures a significant enrichment of screen hits could be shown.

Only for the two smallest stringency settings (GeneScore $\geq$ 13 or 25) for the bCDP was the observed enrichment below the expected one considering the more conservative expectation within the used input data. Above that stringency setting an enrichment of up to three-fold (input data) was statistically confirmed. A total of five screen hits could be identified among the 41 candidates (without AngioTestGroup genes) at the highest stringency setting.

For the gaCDP enrichment of screen hits could be shown for the lowest stringency setting (GeneScore $\geq$ 50). Enrichment of up to three-fold (input data) was statistically confirmed. A total of four screen hits could be identified among the 22 candidate genes at the highest stringency setting (GeneScore = 100).

For the igCDP again an enrichment of screen hits could be shown for all candidate genes of the lowest stringency setting (occurring in at least one AngioProfile). Enrichment of up to three-fold (input data) was statistically confirmed. A total of seven screen hits could be identified among the 52 candidate genes (without UniGene clusters from the AngioTestGroup and the IndicatorGeneSet) that occurred at least in eight AngioProfiles. This estimated enrichment of four-fold (input data) increased even further to eight-fold for candidate genes occurring in at least twelve AngioProfiles.

3.4.3 Procedure Validation - Literature

Presence of additional genes known to modulate angiogenesis served as second means of procedure validation. Therefore a literature search was performed for the remaining candidate genes at high stringency settings (without screen hits, IndicatorGeneSet and AngioTestGroup genes). From the remaining 36 candidate genes of the bCDP with a GeneScore of 100 additional five (13.8 %) UniGene clusters were already known to modulate angiogenesis. From the remaining 18 candidate genes of the gaCDP with a GeneScore of 100 an additional three (16.6 %) UniGene clusters were already known to modulate angiogenesis. For the remaining 45 candidate genes of the igCDP that occurred in at least eight AngioProfiles additional six (13.3 %) UniGene clusters were already known to modulate angiogenesis. Neither the Gene Ontology entries nor the GRIFs of those genes indicated the modulation of angiogenesis, although its association was already known to the public. For that reason, those genes were not part of the AngioTestGroup. With help of Gene Ontology and GRIF it was possible to get a good amount of the known angiogenesis modulators, nevertheless, their annotation is far from complete. Filling these annotation gaps is still a time consuming and highly desirable task.

This comparison of the resulting candidate gene lists with the results of an experimental high-throughput screen for pro-angiogenic factors as well as already known but not yet annotated (in Gene Ontology and GRIF) angiogenesis modulators, showed the feasibility of the approach. Generally all three Common Denominator Procedures accumulate angiogenesis-associated genes.