Project description

The goal of the “Comprehensive geno-glycomic approach to discover new lung cancer biomarkers” project was to develop a novel platform to identify glycobiomarkers including the search for genetic risk factors (Genomics Track) that play a role in the glycosynthetic pathways of important biomarker proteins in COPD and lung cancer (LC) and reveal the corresponding global glycosylation changes in serum (Glycomis Track)[90, 95, 99]. This two Track approach was implemented parallel, as shown in Figure 1.

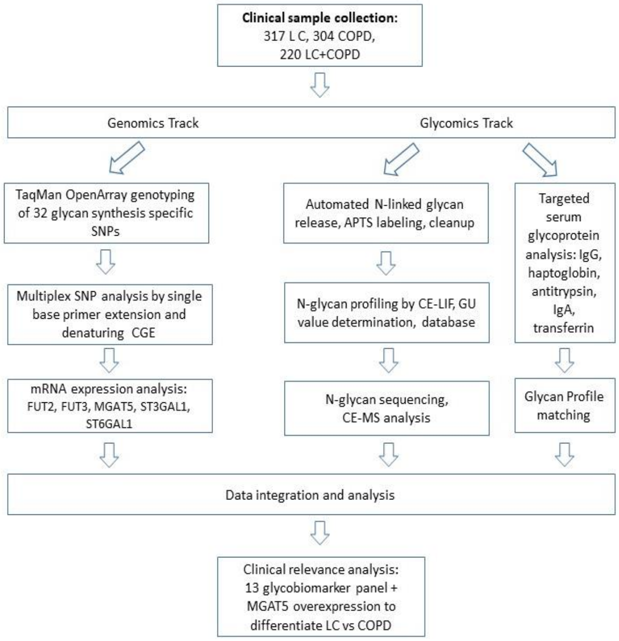
Clinical sample collection:
317 L C, 304 COPD,
220 LC+COPD
Genomics Track
Glycomics Track
TaqMan OpenArray genotyping of 32 glycan synthesis specific SNPs
Automated N-linked glycan release, APTS labeling, cleanup
Targeted serum glycoprotein analysis: IgG, haptoglobin, antitrypsin, IgA, transferrin
Multiplex SNP analysis by single base primer extension and denaturing CGE
N-glycan profiling by CE-LIF, GU value determination, database
mRNA expression analysis: FUT2, FUT3, MGAT5, ST3GAL1, ST6GAL1
N-glycan sequencing, CE-MS analysis
Glycan Profile matching
Data integration and analysis
Clinical relevance analysis: 13 glycobiomarker panel + MGAT5 overexpression to differentiate LC vs COPD

Figure 1. The workflow of the geno-glycomic approach to discover new lung cancer biomarkers.

First we focused on obtaining information about the single nucleotide polymorphic variants (SNP) of glycosyltransferases and glycosidases shedding light on the genetic background of glycosylation changes on several key markers. This endeavor was in full synergy with the recent major focuses of biomedical sciences targeting the carbohydrate modifications of glycoproteins in complex biological systems. Although, genome wide association studies of lung cancer and COPD have already revealed numerous risk factors, it did not provide information on the genetic polymorphisms in glycosyltransferases / glycosidases and the associated changes in N-glycosylation. Therefore, it was necessary to characterize these polymorphisms and the associated serum glycome as completely as possible including linkage and positional information.

In addition, a variety of N-glycans can be possibly present at a particular site (microheterogeneity) making comprehensive characterization even more challenging, especially at their concentration levels in biological systems. The major phases of our work-plan included a Genomics and a Glycomics track. The Genomics Track comprised the SNP-analysis of the enzymes playing significant roles in glycan biosynthesis. Single nucleotide polymorphisms possessing putative regulatory function in genes, which were selected by *in silico* methods using the 1000 Genome Project, HapMap Project, NCBI dbSNP, PolymiRTS and Transfac databases. Genotype analysis was followed by a case–control study evaluated by chi-square-statistics. The Glycomics Track comprised the development and integration of a comprehensive complex carbohydrate characterization platform including temperature addusted denaturation to prevent precipitation, endoglycosidase digestion, glycoaffinity partitioning, fluorophore labeling, capillary electrophoresis – laser induced fluorescent analysis (CE-LIF), generation and utilization of GU databases and capillary electrophoresis - mass spectrometry (CESI-MS) measurements. Exoglycosidase matrix based glycan sequencing was also used to validate GU value based structural elucidation. The generated data was used in our clinical relevance exercise to distinguish between lung cancer (adenocarcinoma) and lung inflammation (COPD) based on a large number of clinical samples by identifying potential genomic and glycomic biomarkers.

Sample collection was finished by the end of the third project year, resulting in a total of 317 lung cancer (LC, stage 3B and 4), 304 chronic obstructive pulmonary disease (COPD, stage Gold

C and D) and 220 LC+COPD patient serum samples. Please note that due to the lower number of patients with LC+COPD comorbidity, the sample collection was stopped at 220 for that cohort to be able to finish the large scale genomic and glycomic studies during the fourth project year. Along with the sample collection process, all buccal samples (930) for SNP analysis were also collected by the end of the third project year.

Description of the results

GENOMICS TRACK: Identification of putative biomarkers was initiated during the first year by in silico SNP analysis and selection of the genes of interest. We aimed to include functional polymorphisms with potential biological effects, thus the following criteria were used applying the dbSNP of NCBI and the 1000 Genomes databases. Matching criteria was based on minor allele frequency values observed in European (EUR) population filtered by the Ensembl Genom Browser. Polymorphisms localized in the regulatory regions assumed to be genetic risk factors for complex disorders, as these variants can contribute to the fine tuning of the regulation of protein biosynthesis, albeit, they usually do not cause significant loss in function. Consequently, polymorphisms in the 5' and 3' regions of the genes were selected. SNPs in the 5' region of the genes might play a role in the modification of transcription factor binding. The matching sequences were than manually searched in the Transfac gene regulation database for identification of single nucleotide polymorphisms located in the transcription factor binding sites of the promoter regions. 3' UTR serves as the binding site for micro-RNAs, the miR-Walk and the PolymiRTS databases were employed to analyze this putative effect of the selected loci. Moreover, polymorphisms with minor allele frequency (MAF) of at least 5% were included. Taking these criteria into consideration, in the first round the following SNPs were selected in the following genes of interest: FUT2, FUT3, MGAT5, ST3GAL1 and ST6GAL1 for analysis of rs2548457, rs2251034, rs632111, rs2306969, rs13406707, rs34944508, rs113932819, rs4736675, rs28687047, rs2922471, rs7839535, rs36118972, rs28366037, rs9875079, rs9716, rs1042757. Multiplex genotyping method based on primer extension and multicapillary gel electrophoresis analysis were first evaluated for high-throughput SNP-analysis. Validation of the technique and the multicapillary electrophoresis system was carried out by analyzing the 5 SNPs (rs4689388, rs4273545, rs1046320, rs1046322, rs9457) in the WFS1 gene [36, numbers represent the entries in the associated Publications section]. It was shown that our system was robust and capable for reliable and efficient genotype analysis using <10 ng genomic DNA. Elaboration of primer design for amplification and extension reactions as well as the optimization of single and multiplex reactions were also carried out.

Based on the technical pilot study [36], multiplex capillary-electrophoresis based genotyping protocols have been elaborated to analyze the relevant SNPs of the ST3GAL1 and ST6GAL1 gene loci. To assure high efficiency and reliability, the first PCRs were carried out in two separate reactions as 3-plex and 2-plex PCR, whereas, primer extension and subsequent capillary electrophoresis analysis were completed in a single 5-plex format to obtain sensitive genotype determination. To increase the reliability of the method, both potential extension primers (forward and reverse) have been designed and tested in the optimization phase. The primers with higher specificity were employed in the subsequent analyses. For verification purposes, allele-specific amplification was also designed for two SNPs to confirm the results of the capillary electrophoresis based method. However, due to the foreseen extra workload of the large scale analysis of the collected 930 DNA-samples for 32 SNPs, we have decided to use the high throughput, but more expensive backup option employing our TaqMan OpenArray system.

DNA purification was carried out using standard protocols including cell lysis in Proteinase K buffer solution and protein precipitation by saturated NaCl, followed by the precipitation of the genomic DNA using isopropanol and ethanol. Sample concentrations were measured by a NanoDrop spectrophotometer and samples with less than 20 ng/µl DNA were excluded from the study. The high-throughput TaqMan OpenArray platform was used for SNP genotyping as this technique allowed completion of real-time PCR and TaqMan probe based genotype analysis in a low-density array format from 33 nL reaction volumes. Target specific primers and probes were immobilized in the through holes of the OpenArray slides, and an automated liquid handling system distributed the genomic DNA samples along with the reaction mixtures onto the slides. In our setup 32 SNPs of 96 samples were analyzed in each OpenArray slide, and 10 slides were used to analyze the entire set of 930 samples. The remaining 30 sample positions were used for one negative, and two positive controls (on duplicate measurement on the same slide and one parallel measurement on two different slides) on each slide, respectively. The complete list of SNPs involved in the study was as follows: rs632111, rs418821, rs601338, rs602662, rs3894326, rs778986, rs1468906, rs28362834, rs812936, rs11673407, rs2306969, rs34944508, rs79594066, rs2922471, rs4736675, rs35166820, rs11782689, rs16904924, rs2142306, rs4736674, rs679574, rs1042757, rs1042642, rs2922467, rs1801380, rs2239611, rs16982241, rs2284750, rs7559, rs28366038, rs9814673, rs2230908. 551 duplicates were technically successful (i.e., both of the two parallel measurements provided reliable genotype information), in case of 536 data points the two genotypes matched with more than 97% reproducibility. The 15 doubtful pairs were omitted from the downstream analysis, although in 11 cases, identification of the incorrectly called genotypes were possible by manual analysis. A subset of 48 randomly selected samples were re-genotyped by the orthogonal (thus independent) primer extension method in conjunction with multicapillary gel electrophoresis analysis. This experiment resulted in approximately the same (97.9%) reproducibility. The call rate of all SNPs was higher than 73%, that of 30 SNPs (93.75%) exceeded 90% and 26 loci (81.25%) could be determined in more than 95% of the DNA samples. 91.16% of the DNA samples could be determined for at least 30 SNPs out of the 32 loci, whereas 47.99% of the DNA samples resulted in reliable genotype in case of all 32 polymorphisms. Since the obtained data demonstrated the high robustness of the genotyping platform, Hardy–Weinberg equilibrium was assessed in case of each polymorphism. A significant discrepancy ($p < 0.1$) between the obtained and expected genotype distribution could be observed in case of six loci, two of them could be unambiguously resolved by manually resetting the genotype clusters, which were incorrectly determined by the automatic algorithm. Subsequent in-depth analysis of the data points shed light on the reason of the bias at the remaining four cases. Our results also suggested that the rs34944508 SNP might modulate the risk for lung cancer by influencing the expression of MGAT5, a typical cancer-associated glycosyltransferase [100]. Please note, that this enzyme catalyzes the addition of N-acetylglucosamine (GlcNAc) in beta 1-6 linkage to the alpha-linked mannose of biantennary N-linked oligosaccharides, thus, increases branching that characteristic for invasive malignancies, and was independently validated at glycomics level [90].

Messenger RNA sample collection and the consequent biobanking started earlier than planned. 18 patients (1 healthy control, 10 chemotherapy and 7 post-operational patients) have been monitored and their blood samples were collected for mRNA and DNA analysis. Gene expression levels of the following 5 genes of FUT2, FUT3, MGAT5, ST3GAL1 and ST6GAL1 were assessed, and RPLPO, GAPDH and ACTB were used as endogenous controls. Expression of the FUT2 gene could not be detected in the samples. Although amplicons of FUT3 could be amplified, the signals were rather low, thus it was excluded from the analysis, as reliable evaluation of the data was not feasible. RNA-isolation was carried out using PAXgene Blood RNA Kit. Concentration of the samples was measured by NanoDrop spectrophotometry, samples with < 20 ng/µl RNA were excluded from downstream processing. Quality of the samples were assessed by determining the $OD_{260/280}$ and $OD_{260/280}$ ratios, and each sample was analyzed by agarose gel electrophoresis to confirm RNA-integrity. The SuperScript™ VILO™ cDNA Synthesis Kit was used to generate first-strand cDNA. The obtained cDNA samples were than further analyzed by TaqMan real-time PCR, and gene expression levels were calculated using the standard $\Delta\Delta C_T$-method. Data analysis was carried out using two different approaches: 1) The average gene expression levels of the investigated target genes were compared in the 4 (COPD, adenocarcinoma, comorbid COPD + adenocarcinoma and control) groups. The ANOVA analysis did not show any significant difference between the gene expression values in the investigated patients. 2) A longitudinal analysis was carried out in case of 5 patients with lung adenocarcinoma, as during the project four or more blood samples were collected form them at different points of their treatment protocol. In these cases, major changes were also observed in the expression level of the MGAT5 gene, interestingly correlating with chemotherapy as well.

GLYCOMICS TRACK: We started this part of the project with the development of new precipitation free denaturation, evaluated deglycosylation inhibition and evaporative labeling methods with the aim to analyze the very complex serum samples and prevent the loss of silalylated glycan structures [59, 60]. The oligosaccharides were released by PNGase F enzyme using optimized reaction conditions (temperature, digestion time) after the removal of the blood sugar content [81]. The release process was validated on the control samples and proved to be useable for the analysis of the collected patient samples using our new coated capillary columns [30, 54]. During the second year, our newly developed glycan tagging protocol was further optimized utilizing an open vial fluorophore labeling approach [73]. Within this part of the work, various labeling conditions were evaluated to obtain the highest tagging efficiency, while minimizing the loss of sialylated structures. The reaction conditions and the protein/enzyme ratio were both optimized for N-glycan release via PNGase F digestion. Considering the high number of samples to be analyzed in the framework of this grant (~1200), at this stage of the method development efforts we also evaluated various automated protocols to ensure reliable and reproducible sample preparation in large-scale. Immobilized recombinant glutathione-S-transferase (GST) tagged PNGase F enzyme microcolumns were applied for rapid and efficient removal of N-linked carbohydrates from glycoproteins [34] that were readily used in automated sample preparation systems, such as with our Biomek 3000 liquid handling robot. Furthermore, we immobilized GST tagged PNGase F onto the surface of magnetic microparticles as an alternative method for rapid N-glycan removal from glycoproteins [29]. Both approaches seemed to be appropriate for automated large-scale sample preparation, that also required novel separation matrices and parameters and high throughput injection methods [5, 27, 32, 33, 37, 57, 92, 94, 98].

At the beginning of the project, we also started evaluating the retrieval process for five glycoproteins of potential biomarker interest (IgG, haptoglobin, human transferrin, human α1-antitrypsin and IgA) from human serum. First we evaluated the Agilent Multi-affinity Removal Cartridge in reverse operation mode, i.e., to recover the removed (i.e., bound) glycoproteins from the stationary phase. In other words, our interest was the opposite that of the column was designed for, thus, required early feasibility studies. Commercially available individual antigens and their mixture were used during the process. Glycan partitioning techniques using magnetic beads were also implemented using crowding technology with high acetonitrile concentration, representing another important element of the advancement of the project [31]. Several other techniques were also evaluated to extract haptoglobin, antitrypsin, IgA and transferrin form their standard mixtures and from spiked plasma samples. We tested the following options: polyglutaraldehyde particles, G-10 and C-18 columns, size exclusion chromatography as well as Agilent HSA/IgG and Multiaffinity Removal spin cartridges. The G10 and C18 columns were not optimal due the high elution volumes required and poor protein recovery. Size exclusion chromatography gave us good protein yield, but it was time-consuming and also necessitated high elution volumes [88]. Reducing the sample volume in a high throughput manner would have been expensive and not feasible to process the ~1200 clinical samples. In addition, with the attempt of the Agilent HSA/IgG and Multiaffinity Removal spin cartridges, the provided elution buffer contained a large amount of salt, representing a problem during the enzymatic N-glycan removal. Based on these discouraging results, we choose the analysis option of the plasma samples at the global N-glycome level and obtained favorable results [82]. Actually, we have discovered that the N-glycans from IgG, haptoglobin, antitrypsin, IgA, and transferrin represent the largest proportion of the human plasma N-glycome, thus should be analyzed as such. Therefore, we considered that the N-glycosylation modifications of the aforementioned proteins can be adequately monitored at the plasma N-glycome level [90]. In addition, total plasma analysis was cost effective and easily integratable into our automated sample preparation platform, but was only used for validation purposes for the mainstream glycomics track [101]. We also compared the pooled serum N-glycome analysis results with the mixture of protein standards of lung disease interest, containing IgG, haptoglobin, antitrypsin, IgA, and transferrin with the %-values corresponding to their normal serum levels. Based on our results, as a first approximation, we concluded that no immunoprecipitation was necessary to gain the required information about the N-glycosylation changes of these glycoproteins as planned to investigate, thus, serum level analysis will be used for all individual sample analyses [82].

In addition to full glycome profiling, deeper investigation was necessary to analyze and identify all carbohydrate structures [74]. To address this issue, an automated exoglycosidase digestion based carbohydrate sequencing system was developed utilizing the temperature control feature of the sample storage compartment of our PA 800 Plus capillary electrophoresis instrument that was otherwise used for the CE analysis of the resulting sequencing fragments [58]. Careful optimization of the reaction temperature, enzyme concentrations and incubation times resulted in effective and fully automated exoglycosidase digestion based carbohydrate sequencing in 60 minutes. To decipher sialylation linkage differences, we used special sialidases, i.e., first to digest the α(2-3) bonds, followed by the α(2-6) level. To increase the precision of our separation endeavor and assure the applicability of the GU database, we introduced a co-injected triple internal standard based instant glycan structure assignment method for the serum N-glycome analysis [28]. This revolutionary approach did not require an accompanying maltooligosaccharide ladder run for glucose unit (GU) calculation, representing significant time saving for the large-scale sample processing we needed. During our profiling and sequencing work, glucose unit (GU) values were primarily utilized for structural elucidation, based on our continuously improved in-house made software tool and the associated publicly available database (GUcal.hu) [74, 93].

During the third year of the project we finished the design and implementation of the comprehensive glycan characterization platform, ready for automated large scale clinical sample analysis [55, 71, 78, 80, 89, 94] including the robust and highly effective evaporative labeling protocol to improve the efficiency of fluorophore glycan labeling [73]. This was accomplished along with a fundamentally novel approach for high sensitivity fluorescence detection in a wide sample concentration range [56] that enabled post-task re-processing of under- or overloaded separation traces without reanalyzing the sample. Quantification of the glycobiomarker candidates represented one of the challenges during our preliminary studies, therefore, a quantitative methodology was developed to address this issue [79]. A numerical process was also proposed to assess the profile and composition similarity of N-glycosylation profiles [70, 77, 87] to differentiate between disease classes (LC, COPD, LC+COPD) and the control. In ambiguous structural elucidation cases, the CE system was connected to mass spectrometry via a porous sheath-less sprayer (CESI-MS) to gain extra information about the glycan structures of interest [97]. The shape of the flow profile played an important role on the efficiency of the resulting peaks in positive pressure assisted reverse EOF based CESI-MS separation mode [75]. Since simultaneous optical and MS detection was not an option with our CESI-MS system, an imaging LIF detector was built and implemented collecting signal in the Taylor cone of the electrospray, i.e., immediately before the sample components entered the orifice of the MS unit [96], a major advancement in the field. The large number of collected patient serum samples were processed on our automated liquid handling platform (Beckman Biomek) by applying the optimized sample preparation method with the processing time of ~3.5 hours for each 96 well plate. For the large-scale analytical work, a multicapillary (12 channel) electrophoresis unit was acquired (C100HT), capable of processing hundreds of analyses per day, necessary to run all samples in triplicates in a reasonable time scale [91]. In addition to preparing the large number of serum samples for the analysis, antigen retrieval and glycan analysis of formalin fixed, paraffin embedded lung tissue-biopsy specimens were also worked out [23, 51, 71, 89].

During the last project year, the N-glycosylation profiles of pooled patient samples of chronic inflammatory (COPD) and malignant (LC) pulmonary diseases as well as their comorbidity (COPD with LC) were quantitatively studied and compared to healthy controls. Sixty-one N-glycan structures were identified in the control human serum samples and since no other glycans appeared in any of the three disease categories, only these 61 structures were quantitatively monitored. At the end of the project, the clinical relevance of the results was evaluated with the leading pulmonology clinicians of the study (Eszter Csanky, MD, PhD and Dr Miklos Szabo) and suggested that certain serum N-glycans could be used as potential markers for the different types of pulmonary diseases [95, 99]. Based on this evaluation, a panel of 13 glycans were identified to adequately differentiate lung cancer, COPD and their comorbidity from the control and LC from COPD in conjunction with some genomic markers (e.g., MGAT5). This was especially applicable for the highly branched sialylated structures as our recent genotyping data suggested the significant increase of MGAT5 activity (results of the Genomics track), i.e., increased branching in lung cancer [90,100], thus should be part of the geno-glycomic biomarker set. In addition, alterations in the N-glycan subclasses, such as fucosylated, mono-, bi-, tri- and tetra-sialylo, as well as mono-, bi-, tri- and tetra-antennary glycans could also carry interesting diagnostic information [90]. The panel of 13 N-glycans along with the information of MGAT5 activity and the corresponding subclasses may provide even more reliable information as they represent the sum of multiple structural changes caused by a given disease.

SUMMARY

The goal of the research project was to develop a comprehensive geno-glycomic approach to identify and correlate genetic risk factors that play roles in the glycosynthetic pathways in COPD and lung cancer possibly causing global N-glycomics changes and set up a glycobiomarker panel with the associated gene expression information. Understanding the effect of the genetic variants (local SNP level and global mRNA expression profiling) of glycosyltransferases and glycosidases shed light on the genetic factors of N-glycosylation changes in chronic and malignant lung diseases. On the genomics side we used commercially available kits. For the deep N-glycomic study, we developed new tools by integrating precipitation free temperature gradient denaturation, inhibition free endoglycosidase digestion, glycoaffinity partitioning, capillary electrophoresis - laser induced fluorescence analysis as well as mass spectrometry for global N-glycomics analysis. All major glycoforms >1% were characterized and the significant N-glycosylation changes (>33%) were considered as biomarkers. After validation we applied the platform to the analysis of the collected human plasma samples of lung cancer (LC), chronic inflammatory lung disease (COPD) and co-morbidity LC+COPD patients. We established a panel of 13 glycans capable to distinguish between malignant and inflammatory lung diseases when used together with MGAT5 expression level information. A pilot analysis compared the global N-glycosylation changes in FFPE lung tissues and plasma.

We particularly accomplished the following:

1) Collected and analyzed a total of 317 lung cancer (LC, stage 3B and 4), 304 chronic obstructive pulmonary disease (COPD, stage Gold C and D) and 220 LC+COPD patient serum samples.
2) Did local and global genetic/genomic studies and identified important SNPs form 930 blood samples.
3) The MGAT5 gene was identified as a genetic factor of increased branching in the lung cancer N-Glycome. Expression level of this gene in blood samples also seemed to be in relationship with the treatment / stages of LC adenocarcinoma.
4) Developed novel precipitation free sample preparation protocols for human serum samples for deep N-glycomic analysis.
5) Introduced new bioseparation methods, separation matrices and analysis protocols based on the physical-chemical properties of the solute molecules (e.g., activation energy requirement).
6) The clinical relevance of the results was evaluated with the leading pulmonology clinicians of the study (Eszter Csanky, MD, PhD and Miklos Szabo, MD).
7) A panel of 13 N-glycan was established that in conjunction with MGAT5 gene expression data apparently differentiated between LC, COPD and LC+COPD as well as all of them from healthy control samples.

Public dissemination:

- Publications: 40 (ΣIF=165.268)
- Submitted publications under review: 8 (IF: CMM (82): 2.196, OH (99): 0.564, JCB (90): 2.813, TrAC(87): 8.428, CS (94): 9.556, JTH (101): 4.662, ACA (88): 5.256, COSB (93): 7.052; ΣIF=40.527)
- Lectures: 77
- Posters: 90.