Fundació La Marató de TV3
21st SYMPOSIUM
Heart diseases

WHOLE-GENOME DIAGNOSIS OF CORONARY ARTERY DISEASE

Josep Comín Colet
Institut d'Investigació Biomèdica de Bellvitge
Roderic Guigó Serra
Centre de Regulació Genòmica

1. Project summary

Coronary artery disease (CAD) is one of the leading causes of death and source of disability in developed countries. In Catalonia, there are more than 34,000 new diagnoses of CAD every year. Health costs for this disease are almost a third of the entire public health budget. Despite the great advances in treatments, due to the increase in the age of the population and other external factors that increase cardiovascular risk, the number of patients is increasing every year. Many of the patients with CAD will eventually develop heart failure. This could be prevented in many cases by earlier diagnosis. An early diagnosis of CAD would allow more intensive management and avoid the transition to more complicated and advanced stages of the disease. The detection of CAD with efficient diagnostic strategies is essential in order to achieve two priority goals: a) diagnosis in patients with suspected CAD but without any acute coronary event; and b) rapid diagnosis in populations of patients with subclinical CAD and with associated risk factors. Early detection will result in rapid intervention, which will allow us to significantly reduce these massive costs that the disease causes in both lives and resources. Therefore, advances in the methods of diagnosing CAD are needed.

Due to their easy detection in whole blood samples, expression of RNA profiles has become a promising technique in the establishment of potential biomarkers. The set of RNA molecules in a biological sample represents a dynamic picture of the cellular state at that specific moment. Blood biomarker analysis has focused on a small fraction of our genome sequence, the 2% that encodes for protein. The remaining 98%, with thousands of non-coding RNA species, is yet to be explored. Thanks to advances in new sequencing technologies, it is now possible to obtain the profile of the entire human genome in a single measure. This method is known as RNA sequencing (RNA-Seq), and detects both coding and non-coding RNA, which improves genetic signature and candidacy of potential biomarkers. Leveraging this technique, the GTEx Project (Genotype Tissue Expression) has traced entire genome RNA expression to hundreds of patients with a complete medical history, including patients with CAD.

The current project aims to develop a diagnostic score to evaluate and diagnose the presence of CAD, based on an algorithm that incorporates both clinical and genetic expression data of the patient. To this end, the definition of a set of biomarkers for

gene expression is needed to provide information about the differential expression between cases with CAD and controls without CAD. The development of these diagnostic models, according to international recommendations, requires cohorts of patients for both the derivation of the model and the validation of the model.
The methodological phases of the study are:

1. Identification of potential CAD-associated RNA biomarkers in a retrospective cohort:
a. Definition in the first study cohort (first retrospective cohort of GTEx) by bioinformatics analysis of the data from RNA-Seq, of a broad set of biomarkers (biomarker set #1), that show the differential expression between cases and controls (with statistical association.
b. Definition by clinical, methodological, statistical and previous knowledge of a reduced set of RNA biomarkers from this first analysis (biomarker set #2).

2. Confirmation of the RNA biomarker set (biomarker set #2) in a new cohort of patients (validation cohort).

3. Refinement of the biomarker set #2 into a biomarker set #3: definition of a set of biomarkers based on this data, smaller and better discriminating between cases and controls (biomarker set #3), and with this set define a diagnostic algorithm that also integrates clinical data of the patient (expressed in the form of a score).

4. Validation of the algorithm in new prospectively evaluated patient sets.
Therefore, this study has involved a retrospective analysis of the differential genetic expression between patients with and without CAD from the GTEx cohorts, and the prospective recruitment of a Catalonia-wide patient cohort that, for various reasons, were undergoing non-invasive evaluation of the presence or absence of CAD by means of coronary computed tomography (CT). We originally planned a derivation of the biomarker sets #1 and #2 from the GTEx (retrospective) cohort, and then a validation of the biomarker set #2 and refinement of this latter into a biomarker set #3 in the second (prospective) cohort in order to build a final score that integrates genetic and clinical data. This score will make a non-invasive diagnostic approach to the presence or absence of CAD in other populations.

This work seeks to find for the first time new blood RNA biomarkers indicative of CAD, using the data provided by the GTEx Project. The idea is to generate an algorithm to diagnose CAD more precisely than the currently available scoring systems. This would be the basis of an affordable non-invasive medical diagnostic test for patients with suspected CAD.

2. Results

The complete project has recruited a total of 353 patients. Data obtained from them include: baseline clinical data, cardiac imaging (CT), vital and clinical status at follow-up and laboratory samples. Special care has been placed in obtaining high quality data in terms of clinical information, imaging and blood samples.

Clinical Outcomes (Cohorts recruited)

The mean age of the patients was 60 ± 11 years and 54% were women. The origin of the patients was 87.5% European.

The anthropomorphic variables collected were as follows: mean weight 76.7 ± 15 kg, height 162.7 ± 20 cm, waist circumference 96 ± 14 cm, and hip circumference 105 ± 10 cm. The vital signs at the time of testing were: HR: 71 ± 14 bpm, systolic blood pressure 131 ± 23 mmHg, and diastolic blood pressure 74 ± 14 mmHg. The presence of risk factors, other relevant pathologies and medical therapy are summarized in the following table.

Risk factors and relevant pathologies	**Frequency**	**%**
Arterial hypertension	193	55
Diabetes mellitus	55	16
Dyslipidaemia	181	51
Active smokers	54	15
Ex-smokers	104	30
Family history of early HF	30	8.5
Medical therapy	**Frequency**	**%**
Beta blockers	131	37
Diuretics	80	23
ACEI/ARBs	147	42
Calcium antagonists	46	13
Insulin	14	4
Metformin	43	12
Statins	164	46
Ezetimibe	26	7
Aspirin	152	43
Clopidogrel	19	3

In most patients (74%) coronary CT was performed as a part of chest pain diagnostic workup. Chest pain was classified as typical (16%), atypical (55%) or non-coronary (14%). Calcium was scored in 82% of patients, and the mean of the global population was 114 ± 246 Agatston units. With regard to the analytical values of the recruited population, we highlight the following parameters: Hb 14.4 ± 6 gl / dL, creatinine 1.3 ± 5.5 mg / dL, Hb A1c 5.8 ± 2%, total cholesterol / LDL / HDL was 186 ± 40 mg / dL / 106 ± 44mg / dL / 54 ± 16 and triglycerides 122 ± 66ng / dL.

As expected, in the analysis of the coronary CT anatomy we found that 75% of patients did not have CAD. From the patients studied with CAD we observed that 49 (14%) had significant disease of one vessel, 29 (8%) had CAD of two vessels and 10 (3%) had CAD of three vessels.

In terms of follow-up, the survival of the population for the year was 94% (3 patients died). About 5% of patients required revascularization during follow-up.

Genomic Analysis Process and Results

In our project, from the GTEx database we have been able to carry out a complete genome analysis that has allowed us to define the differential genetic expression between cases of CAD and controls without the limitations of searching only what had previously been published in the literature or based on candidate genes. Through a multidisciplinary work process integrating clinicians, biologists and bioinformatics we have defined an initial set of 313 genetic biomarkers that are differentially expressed between these subjects with and without CAD. This biomarker set constitutes the biomarker set 1 (biomarker set #1). Based on the literature review and the clinical interpretation of the initial results, the research team has decided to define a smaller set of expressed genes up to 112, for validation in subsequent cohorts. This set of genes constitutes the next set of biomarkers (biomarker set #2).

At this point, and before starting any validation of the biomarker set #2 in the prospective cohort recruited for this study, we decided to perform a quality control of the data obtained from the GTEx analysis, comparing it with our samples by direct sequencing. The arguments in favour of this parallel validation were:

1. The low consistency and reproducibility of the findings in several independent studies to identify genetic panels associated with heart disease.

2. Recent advances in RNA-Seq technologies in relation to the methodology used in GTEx.

3. The overlap between the genes identified in different studies is not perfect.

4. The genetic and socio-cultural substrate of the individuals of the GTEx project is very different from that of the population of Catalonia that we are going to analyse. Therefore, complete RNA-Seq was performed following the same RNA extraction and manipulation protocol performed in the GTEx Project (Paxgene kit). This new sequencing of a subgroup of our own population collected in Catalan hospitals would allow us to:

1. Compare the gene expression profiles of the GTEx population with our Catalan population.

2. Validate whether the genes identified in the biomarker set #1 (extracted from the GTEx population) can be extrapolated to the Catalan population.

3. Refine and restrict the biomarker set #1 with those genes that have been reproduced in both populations, and that appear as markers of CAD in the Catalan population and that are considered interesting.

The selection of patients initially recruited at the beginning of this experiment was selected based on the lesions observed in the CT studies. Three groups were established:

GROUP 1 - no lesions: CT scans showed no vessels affected.
GROUP 2 - intermediate lesions: CT scans showed lesions of a maximum of 49% in 1 or more vessels and / or segments.
GROUP 3 - severe lesions: CT scans showed lesions > 49% in at least 1 or more vessels and / or segments.

Unfortunately, after two complete batches of sequencing in 60 patients (20 patients per group), and the addition of an extra RNA purification step, no analysis could be performed as there were degradation sample problems in the haemoglobin depletion step of the protocol, and therefore we could not obtain an optimal sequenced sample for subsequent bioinformatics analysis.

Therefore, the research team together with the CRG Core laboratory, planned a new attempt of sequencing the samples using an exclusive new in house protocol. This protocol would make it possible to solve the degradation problems, since we were certain that the quality of the samples taken before haemoglobin depletion was high. For the new RNA-Seq analysis with a new haemoglobin depletion kit 10 samples were chosen this time, 4 from group 1 (no lesions) and 6 from group 3 (severe lesions), to refine the new protocol. In this batch the samples did not show the degradation that the previous ones presented at this point in the process. Therefore, we have been able to set up an optimal sequencing protocol that allows us to process the samples properly and then bioinformatically analyze them in order to continue with the validation. The final sequencing data was obtained at the end of December 2019. A preliminary bioinformatics analysis of the 10 samples from both groups, 1 and 3,

indicates that there are more than 200 differentially expressed genes between groups. Most of them are overexpressed in patients, as was initially found in the analysis of GTEx data. This result indicates that our local cohort is partially comparable to the data present in the GTEx.

Currently, we are increasing the number of samples to be sequenced with the new RNA-Seq processing protocol, increasing the number of patients up to about 30 for each study group (groups 1, 2 and 3). Once all the samples have been sequenced and analyzed, we will be able to define the differential RNA expression panel between cases and controls. One of the problems when analyzing the results of RNA-Seq and the possible population differences is the great variability among the patients. To reduce this variability, computer programs are being implemented to eliminate the potential batch effect. The rest of the samples we obtained (total 353 patients) are stored at -80° for future applications foreseen in the final phase of the study (validation cohort).

4. Relevance with the possible future implications

Despite not having the final results of the project, sequencing patient samples from our own population will add value to the whole study as we will be able to validate them more robustly, generate a more accurate diagnostic algorithm, and at the same time compare the results of our population with the data of the first GTEx bioinformatics analysis.

The discovery of new blood biomarkers for CAD, and the ability to define a diagnostic score, would improve and simplify the diagnostic process for CAD (there are tens of thousands of new cases in Catalonia each year). At the same time, early detection would mean rapid treatment, and the evolution of the disease could be stopped, reducing clinical events and the healthcare cost associated with such a prevalent disease.

On the other hand, this new tool could help in the rapid screening of asymptomatic CAD patients, improving stratification of cardiovascular risk. In addition, the algorithm derived from the project could be incorporated into the clinical guidelines, varying and

improving daily medical practice. Finally, it would allow us to describe new pathophysiological mechanisms, and thus to design new therapeutic strategies for CAD. In summary, this project has the potential to make significant advances in clinical diagnosis, resulting in a considerable improvement in the lives of thousands of people in Catalonia suffering from or without diagnosed CAD.

4. Generated scientific bibliography

Due to the protocol changes and problems discussed above, no article has been generated so far; although currently we are working on the manuscript about the study design with the baseline data from patients. The resultant exploitation of the study will be made once the biomarkers discovery phase at least has been completed, and later we will continue with the validation and the possible generation of the diagnosis score.