Slide 1: the leader will introduce us.

Good afternoon,

**Slide 2:**

Polycystic ovary syndrome (PCOS) is a common hormonal disorder that affects a significant number of women worldwide. It is characterized by hormonal imbalances, manifestating with various symptoms and lead to a chain of consequences, such as missed or irregular menstrual periods, excess hair growth, acne, weight gain, insulin resistance, infertility and so on.

The exact causes of PCOS remain unclear, but there are several factors believed to contribute to its development. Genetic factors play a significant role, as women with PCOS often have relatives with the condition. According to the DisGeNET database, 71 polymorphic loci and 153 genes (both protein-coding and regulatory genes) associated with PCOS are known.

However, the complete understanding of the participation of polymorphic loci in lncRNAs genes in the context of their distribution among common human populations has not been revealed.

**Slide 3:**

Reading from the slide about aim and objectives.

**Slide 4:**

This slide provides a graphical representation of our worflow. As a source of significant SNPs associated with PCOS, we used the GWS Catalog resource. 29 GWASs, including 111 polymorphic loci were selected for further studying.

Next, from the main list of SNPs, we mapped each SNP using the UCSC genomic browser, and filteredout two sets of SNPs: the first one, we filtered out just SNPs those were within or close to the lncRNA genes. In the second set, only SNPs that occurred in the GWASs two or more times were included. For these two sets, we applied several types of computational approaches: we studied the prevalence of effect alleles in common populations, examined the expression of mapped genes, and performed a Gene Ontology analysis for the second set of mapped genes.

**Slide 5:**

This slide exhibits the summary of the involded studies, involving 100 170 cases and 2 088 636 controls from four common populations: Europeans, Chinese and Koreans (which is EAS), Africans and Americans.

**Slide 6:**

On the next slide, using UCSC Genome browser and hg19 genome assembly, we determined the list of SNPs that are within the lncRNA genes (represented as green rows in the table). For those SNP that are intergenic, we filtered out only SNPs, that have closest lincRNA gene (yellow rows) and also distances to these genes and gene position (upstream/downstream). The 18 SNPs were filtered out and 17 lncRNA genes were mapped on these first set of SNPs.

**Slide 7:**

On the next slide, using population genetics data from the 1000 Genomes Project Phase 3, we extracted effect allele frequencies for each SNP, clustered them using hierarchical clustering algortihms and plotted a heatmap using R package superheat. It is possible to split the set of lncRNAs SNPs into three clusters based on prevalence values. Top cluster has “low” frequency values from 0.02 (show on the slide) to 0.365 (show on the slide). The second cluster has “moderate” frequency values from 0.2042 (show on the slide) to 0.6311 (show on the slide) and the third cluster has “high” frequency values from 0.484 to 0.914. This analysis allows us to identify the SNPs and the corresponding lncRNA genes that are most widespread in the common human populations (6 SNPs and 7 lncRNA genes). Some SNP (represented in the table on the slide) were exlcuded due to missing frequency values in some or all populations.

**Slide 8:**

Next, we counted in which tissues each lncRNA gene from our list is expressed and built a histogram. It has been shown that 7 out of 17 lncRNA genes are expressed just in testicles. The other 6 lncRNA genes are expressed in testicles and in other tissues. 4 lncRNA are expressed in the cervix and tibial nerve. In the remaining tissues on the slide, 3 lncRNA from our list are expressed.

The reason why the 7 lncRNAs highlighted by us as markers of PCOS pathogenesis are expressed just in testicles remains unclear. Perhaps we do not fully understand the genetic mechanisms of folliculogenesis at the moment and the answer hidden in the interaction of genes in the early stages of embryogenesis.

**Slide 9:**

Next, we used a different approach in determining significant markers of PCOS pathogenesis - we filtered out the SNPs that appear in the GWAS studies two or more times. Next, we mapped them into genes (both lncRNA and protein-coding genes) and the results are presented on the slide. There are 16 SNPs and 15 mapped genes. SNPs appear just in two or three studies (reflected in the last column). With this approach, only 2 SNPs were identified, located within the lncRNA genes.

**Slide 10:**

For the second set of SNPs, we repeated plotting the heatmap of prevalence of the effector alleles in 5 human populations. Three clusters can also be distinguished here, the highest frequency values are found in the lower cluster, represented by 4 SNPs (1 SNP belongs to lncRNA gene IRF1-AS1). This lower cluster has frequency values from 0.4417 to 0.896. 1 SNP was excluded due to missing frequency values.

**Slide 11:**

Since, after studying the expression of genes (which are most protein-coding) from the second set were detected in a variety of tissues, the expression of two lncRNA genes are shown: LINC02905 and IRF-AS1. LINC02905 has highest expression in ovary (about 10 TPM) and also in few tissues (Testis, Stomach, Heart, Coronary artery) and can be considered as a potential biomarker for experimental validation. IRF-AS1 is expressed in almost all 54 tissues, studied by GTEx project.

**Slide 12:**

In addition, for the second list of genes, an analysis of gene ontology was performed using the Panther database. The parameters are indicated on the slide. The result demonstrates the participation of *ERBB4, YAP1* genes in cardical muscle tissue regeneration and the participation of *ERBB4, YAP1, PROX1* genes in the positive regulation of hearth growth.

**Slide 13:**

As a result, based on the studied associations from the GWAS Catalog and applying all of the abovementioned computational methods, it is possible to identify significant loci that can be included in diagnostic test systems and as targets for genome editing systems (CRISPR/Cas9). *(Read the list of SNP and lncRNA genes from the slide)*

**Slide 14:**

I am telling about our team of bioinformaticians (computational biologists).