**Final Report (625356 CLLrisk)**

1. FINAL PUBLISHABLE SUMMARY REPORT

This section normally should not exceed 2 pages.

This is a comprehensive summary overview of results, conclusions and the socio-economic impacts of the project. The publishable report shall be formatted to be printed as a stand-alone paper document. This report should address a wide audience, including the general public. Moreover, do not include in the summary report any confidential information, whose publication might undermine the protection of commercial interests, including intellectual property, or privacy and the integrity of the individuals, in particular in accordance with Community legislation regarding the protection of personal data.

Please ensure that it:

• Is of suitable quality to enable direct publication by the REA or the Commission.

• Is comprehensive, and describes the work carried out to achieve the project's objectives; the main results, conclusions and their potential impact and use and any socio-economic impact of the project. Please mention any target groups such as policy makers or civil society for whom the research could be relevant.

• Includes where appropriate, diagrams or photographs and the project logo, illustrating and promoting the work of the project.

• Provides the address of the project Website (if applicable) as well as relevant contact details.

***Sample collection and sequencing:***

*Cases*

In total 450 exomes of CLL samples (matched normal and tumor) and 150 whole genomes have been sequenced on the Illumina Hiseq2000 sequencing machine.

*Controls*

A total of 780 individuals from our in-house exome sequencing data was used as controls for our study. Some of these individuals belonged to different diseases including Obsessive Compulsive Disorder (OCD), Intellectual Disability, Alopecia Areata, Fibromyalgia, Parkinson, Essential Tremor, Cystic Fibrosis, Spinocerebellar ataxia, Neuromyelitis Optica, Stroke, Ataxia, ChiariMalformation, Myasthenia, Progressive Encephalopathy, Immunodeficiency, and Vitiligo. One individual from this control cohort was centenarian and there were 63 healthy control samples.

***Alignment and variant calling from the sequencing reads:***

Exome sequencing data: All the CLL cases and controls were aligned and variant calling was performed in exactly the same manner. For alignment, the reads generated from the sequencer were aligned to the human genome reference sequence using Burrows-Wheeler alignment (BWA-mem algorithm) (Li and Durbin 2009). Any duplicate reads was flagged using the MarkDuplicates algorithm from Picard (http://picard.sourceforge.net/). Local realignments and base quality recalibrations were performed using GATK haplotype caller (DePristo et al. 2011) and gVCFs (genomic VCFs) were generated for the BAM files for all the samples. The gVCFs were then combined and joint genotyping of single nucleotide variants (SNVs) and indels were performed. The variant quality score recalibration (VQRS) of GATK was utilized to filter variants based on a variety of features including call rate, depth of coverage, minimum allele frequency, fisher strand etc. The SNVs and indels were further annotated with their damage potential predicted from multiple bioinformatics tools, allele frequency information from the 1000 genome (1000 Genomes Project Consortium, 2010) and the Exome Variant Server (EVS) database ((National Heart, Lung, and Blood Institute Exome Sequencing Project, found at http://evs.gs.washington.edu/EVS/). Quality control was performed to remove any sample that behaved as outliers. The germline variants detected in the normal DNA samples from the CLL cases were utilized for further rare variant association analyses.

Whole genome sequencing data: To identify structural variants from the whole genome data, we have used PeSV-Fisher pipeline. It has been run on all the 150 CLL samples and currently the analysis is in progress.

***Rare Variant Association Analyses***

Three different methods for rare variant association analyses were performed: Kernel-based adaptive cluster (KBAC) (Liu and Leal 2010), Optimal Sequence Kernel Association Test (SKAT-O) (Lee et al. 2012) and Mixed Effects Score Test (MiST) (Sun et al. 2013). KBAC is a unidirectional rare variant gene based test whereas SKAT-O and MiST is a linear combination of unidirectional and variance-component tests. To increase sensitivity, we utilized a combination of these three methods for rare variant association tests (α=0.05). Variants with a population minor allele frequency ≤ 0.05 were tested for case-control rare variant enrichment analysis.

We have identified a total of 43 genes that were significantly enriched in cases compared to controls. Some of the pathways represented by these genes are progesterone mediated oocyte maturation, signal transduction, proton pump inhibitor pathway etc. We are currently following up with the 15 candidate genes that might be associated with germline susceptibility in CLL. One of the top ranking gene is in the process of Sanger validation and replication in additional 100 CLL cases.

***Targeted Panel***

We designed a targeted panel (Nimblegen SeqCap EZ library) consisting of coding regions of 358 genes, 945 cancer susceptible SNPs and 166 microRNAs. The relevant genes were selected by extensive literature search based on their potential involvement in causing cancer. This targeted panel was sequenced in a cohort of 95 CLL, and 284 healthy control samples. The set of additional cases was available through collaboration with Hospital del Mar and controls through Multi-caso control-Spain (MCC-Spain; http://mccspain.org/). Sequencing was performed on Illumina HiSeq2000 platform as 125 base pair paired-end reads. Alignment of the sequencing reads and variant calling was performed in a similar way the exome data on CLL was analyzed. As previously discussed, rare variant association tests were performed using a combination of KBAC, SKAT-O and MiST. Only two genes were common between the significant genes from the exome study and this targeted panel. We were able to show gene A to be replicated in these additional CLL cases. Additionally, four cancer susceptible SNPs were found to be significantly associated in our CLL cases than in controls. These SNPs are rs27524, rs757978, rs11083846 and rs7097.

***Role of multiple germline mutations in CLL predisposition***

We hypothesize that in many CLL cases, only one germline mutation may not be sufficient to cause cancer predisposition. In such cases, may be a combination of germline mutations will result in predisposition to cancer. We have utilized a random forest method and identified a total of 15 genes that separate cases with controls. This is with an accuracy parameters of precision error of 20%, sensitivity of 53%, specificity of 94%, positive predictive value of 84% and negative predictive value of 78%.

The results from this study has potential in the study of disease diagnosis in CLL.