Candidate mutations for early-onset lung cancer by family genome sequencing

Vangelis Simeonidis[1,2], Jared Roach[2], Mary Brunkow[2], Gustavo Glusman[2], Sheila Reynolds[2], Rudi Balling[1], Leroy Hood[2], David Galas[2], H.-Erich Wichmann[3], Sara Grimm[4], Richard Gelinas[2]

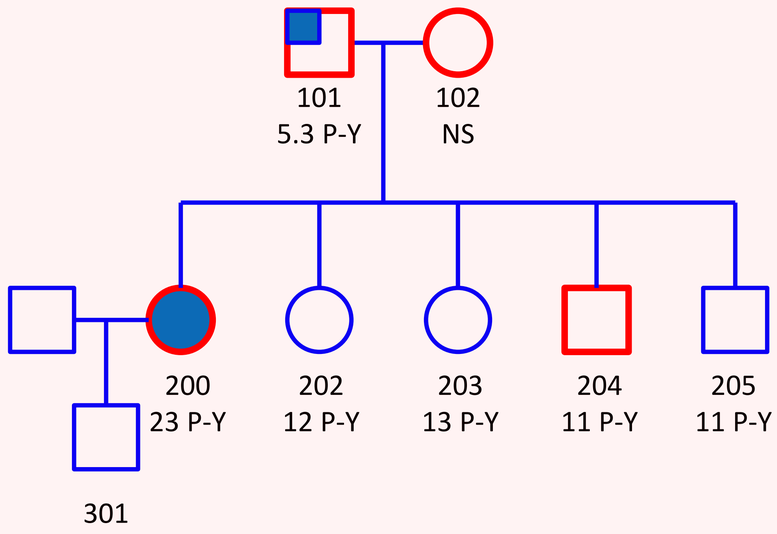
[1]Luxembourg Centre for Systems Biomedicine; [2]Institute for Systems Biology; [3]Helmholtz Centre Munich; [4]National Institute of Environmental Health Sciences

Introduction

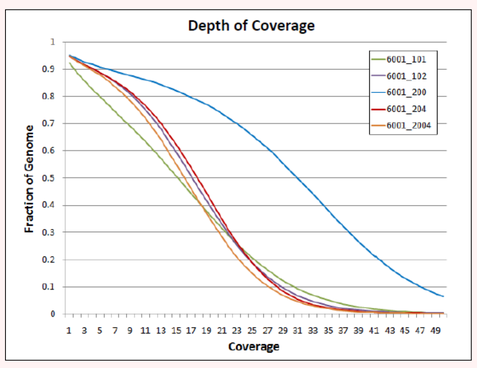
Early-onset lung cancer, often defined as lung cancer in patients less than 50 years of age, has been studied as a rare, but distinct, sub-type of lung cancer. Genome-wide association studies (GWAS) have linked several genes with this form of malignancy. Here, we analyze the whole-genome sequences of four members of a family (two siblings and the parents), one of which has been diagnosed with lung cancer. Family genome analysis enables us to narrow the candidate genes (if any) for this type of cancer, assuming a Mendelian inheritance pattern for early-onset lung cancer. The results of our analysis are discussed and conclusions about possible causative mutations for early-onset lung cancer are drawn, demonstrating the value of sequencing the genomes of a complete family, especially if an inherited disease is suspected.

LUCY study family

We sequenced the genomes of a family quartet in which one of the offspring was diagnosed with early-onset lung cancer. The family, which participated in the Lung Cancer in the Young (LUCY) GWAS study*, has a history of heavy smoking, given in pack-years (P-Y). Individual 200, the proband, was diagnosed with adenocarcinoma of the lung at the age of 48. Her father, 101, had head and neck cancer (sequenced genomes in red).



Whole genome sequencing



The genomes of individuals 101, 102, 200 and 204 were sequenced by ABI-Life Technologies. The genome of 200 was also sequenced by Illumina.

Average coverage for the proband is 29, whereas for the rest of family it is 16. The second sequencing of the proband by Illumina yielded greater coverage.

There are 18 million sequence variants (single nucleotide and indel) in the proband genome and 14 million in her brother's genome.

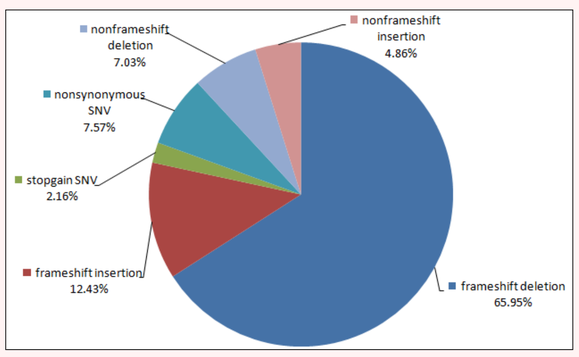
First plan: study sequence variants in exons only, which yields 84K exon variants (reducing the focus to less than 1% of the data). Our objective is to narrow the exonic potential variants in the proband genome to a smaller set of interesting potential variants.

We filter for **exonic variants** that are i) novel; ii) observed with high coverage; iii) not observed in the brother's genome sequence, and; iv) have potentially significant protein-coding effects.

Results

The DNA source was blood, which led us to concentrate our analysis on Mendelian inheritance models. More than 18 million sequence variants were initially identified in the proband through comparison to the hg19 reference genome. We reduced this list to fewer than 200 potentially functional variants (e.g. single nucleotide variations and short indels) present in the genomes of the proband and at least one parent, by applying a series of filters:

- Since known variants are potentially of lesser interest, we set aside entries that are already catalogued in dbSNP.
- The criteria for detecting a "potential variant" are extremely lax and prone to collecting false positives. Therefore, we retained only entries with significant coverage.
- Because one offspring is affected and the other is not, we removed from consideration those potential variants that are observed in both the daughter and son.
- As some variants are expected to have greater impact than others, we retained only entries predicted to result in significant alterations in protein function or folding.



The analysis, after the four filtering steps, yields 185 proband genome variants as potential mutations of interest. We refine the list of candidate mutations further by comparison to gene candidates from GWAS studies and genes that are mutated in lung cancer tissue as recorded by The Cancer Genome Atlas.

Conclusions and future work

We sequenced the genomes of a four member family in which one of the offspring had early-onset lung cancer. The analysis did not give any matches between the 185 identified genome variants and gene candidates from GWAS studies. A limitation of our study was that with no access to tumor tissue, we could not look for acquired or somatic mutations that might be confined to the tumor itself.

To make the inheritance pattern explicit, we will establish the parental origin of the offspring's genomes through phasing of their chromosomes. This will increase the specificity of our analysis and allow us to identify variants that might have been missed in the above approach.

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

*The LUCY study is a multicenter study with 31 participating clinics in Germany, coordinated by the LUCY-Consortium (Head: H.E. Wichmann).
This work is funded by the University of Luxembourg and the LCSB. This presentation was partly made possible by a Travel Fellowship awarded by ISCB with grant funds obtained from the Department of Energy Office of Science, the National Science Foundation Bio-Directorate, and the NIH National Institute of General Medical Sciences.

For more information please contact me at evangelos.simeonidis@uni.lu or on twitter: @vangos