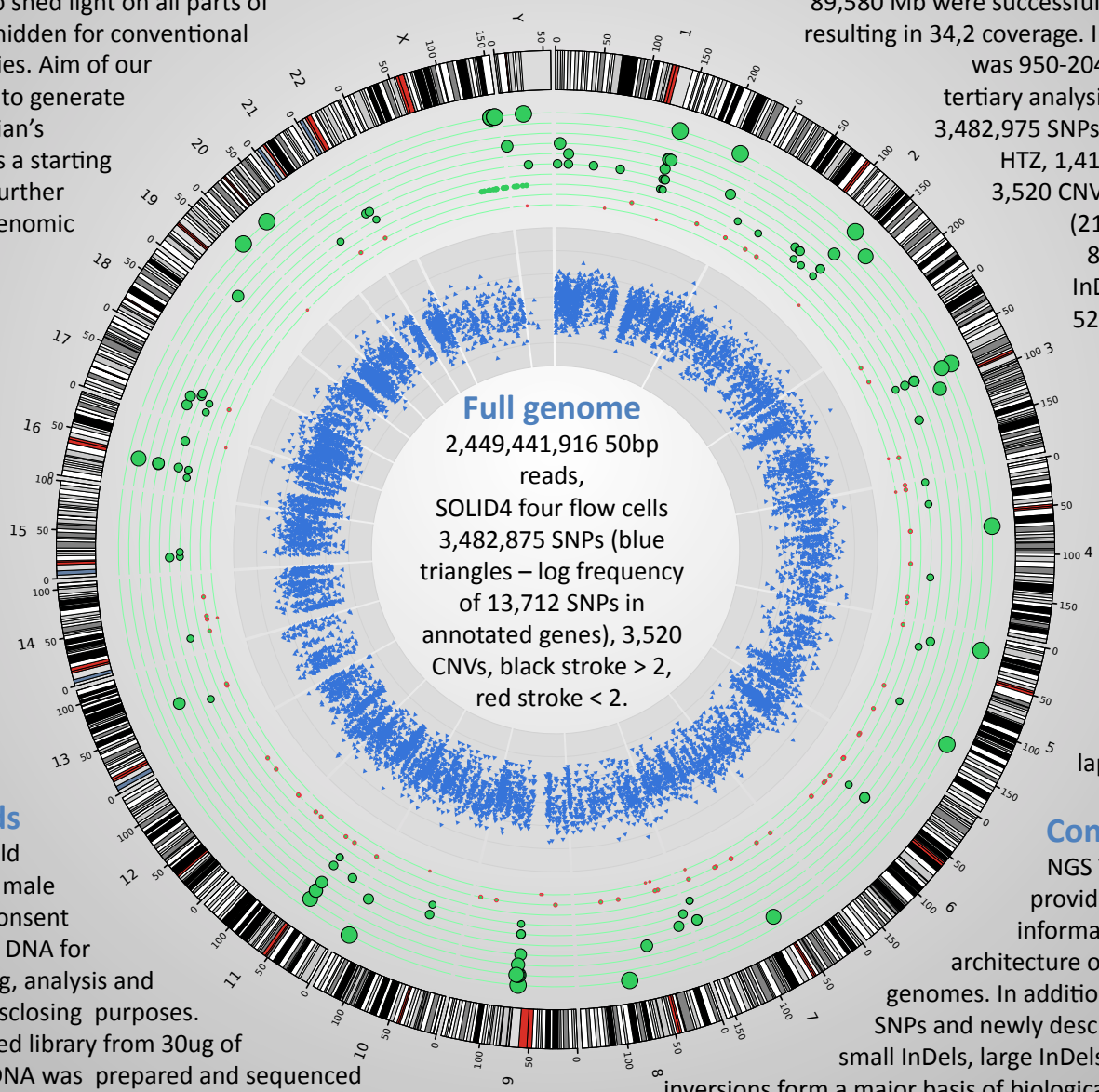
Preliminary Description of the Genome of the Single Individual from Northern Europe

Rutt Lilleoja[1], Aili Sarapik[1], Ene Reimann[2,3], Ülle Jaakma[1], Sulev Kõks[2,4]

[1]Estonian University of Life Sciences, Estonia, Department of Reproductive Biology; [2]University of Tartu, Estonia, Department of Physiology; [3]University of Tartu, Estonia, Department of Dermatology; [4]University of Tartu, Estonia, Centre of Translational Medicine

Introduction and Aim

Next Generation Sequencing enables fast and high throughput sequencing of very complex genomes. For the first time in the history of genetics it is possible to shed light on all parts of genomes hidden for conventional technologies. Aim of our study was to generate first Estonian's genome as a starting point for further national genomic projects.



Methods

39-years old Caucasian male gave full consent to provide DNA for sequencing, analysis and publicly disclosing purposes. Mate-paired library from 30ug of genomic DNA was prepared and sequenced with SOLID3+ & 4. Color space fasta files (.csfasta) and appropriate quality files (.qual) were mapped and paired to the reference genome hg18 version. Mapping (mapreads algorithm) and tertiary analysis was performed using the Bioscope Software (ver 1.3).

Results

Initial sequencing resulted in the 2,449,441,916 50-bp reads. With mismatch penalty -2.00 and clearzone 5, average mapping was 75.6%. Altogether 89,580 Mb were successfully mapped, resulting in 34,2 coverage. Insert range was 950-2040 bp. With tertiary analysis we found 3,482,975 SNPs, 2,067,200 HTZ, 1,415,775 HOZ, 3,520 CNV segments, (218 verified), 87,451 large InDel regions, 52 inversions, 285,864 small InDel segments. 11% are novel SNPs, 34,992 InDels were in genes. From inversions, 19 over-lapped genes

Conclusions

NGS Technology provides in-depth information on the architecture of individual genomes. In addition to known SNPs and newly described SNPs, small InDels, large InDels, CNVs and inversions form a major basis of biological variation. Complex analysis, data processing and visualization are needed for easily accessible personalized genomic data. The European Regional Development Fund together with the Archimedes Foundation and Estonian Science Foundation (grant GARFS7479) supported this study.