Whole-genome sequencing methods for CNV detection

Zuzana Klinovská[1], Marcel Kucharík[1,2], Jaroslav Budiš[1,2,3], Tomáš Szemes[1,2,4]
[1]Geneton Ltd, Bratislava, Slovakia
[2]Comenius University Science Park, Bratislava, Slovakia
[3]Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
[4]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava

The importance of copy number variants (CNVs) in reference to human health is rising continuously. They are linked to various types of disorders, syndromes and recent studies showed they can be used as a biomarker for cancer screening. Next-generation based technologies enabled faster and more cost-effective ways for CNV detection, which could be applicable in practical use. In this article, we describe the general approach of CNV detection tools, which are based on next-generation technologies and where data was obtained by whole-genome-sequencing in particular. We provide a general overview of the whole process of detection of a CNV, provide historical context and briefly describe the individual steps of the process.
Key words: CNV, whole genome sequencing, Detekction of CNV, NIPT

Detekcia CNV metódou celegonómového sekvenovania
Význam CNV v súvislosti s ľudským zdravím neustále rastie. Tieto varianty sa spájajú s rôznymi syndrómami a poruchami a dokonca môžu slúžiť ako biologický označovač pre prítomnosť rakoviny. Technológie NGS umožňujú rýchlejšie a lacnejšie detegovanie týchto variantov, čo zjednodušuje ich použiteľnosť v praktickej medicíne. V našom článku opisujeme všeobecný prístup strojov pre detekciu CNV založených na technológii NGS. Špeciálne sa zameriavame na získavanie dát metódou celogenómovej sekvencie. Poskytujeme všeobecný prehľad celého procesu, históriu vývoja rôznych metód pre detegovanie CNV a tiež sa bližšie vyjadrujeme k jednotlivým krokom samotného detegovania CNV.
Kľúčové slová: CNV, celegonómové sekvenovanie, detekcia CNV, NIPT

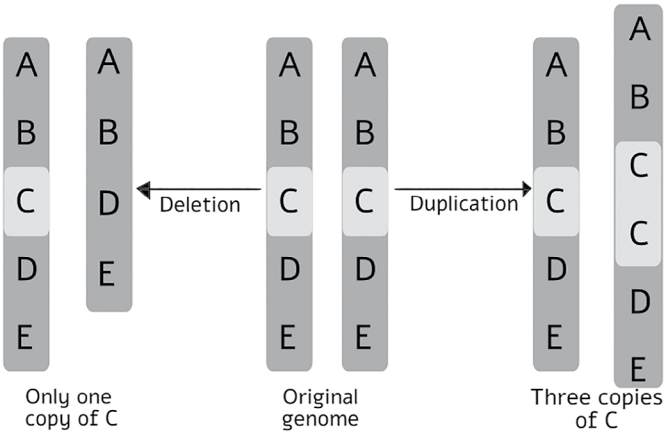
Newslab, 2022; roč. 13 (1): 37 – 41

Introduction to sequencing

The bioinformatic field is mainly interested in the development of methods and software tools that contribute to the understanding of biological data. This interdisciplinary field intervenes with various biological disciplines, especially with genomics, which focuses on the structure, function, evolution, mapping, and editing of genomes. Knowledge from bioinformatic studies and research can then be used in practice to treat various diseases and abnormalities, furthermore, it can also help us understand evolution and our origin.

To obtain biological data, DNA has to be firstly sequenced in order to determine the order of nucleotides. Four nucleotide acids cytosine [C], guanine [G], adenine [A], and thymine [T] make up the whole DNA molecule, computer programs store them in short as A, C, T, G letters. Sequencing itself underwent rapid development throughout the years. The first human genome was sequenced in 2001 and the work took thirteen years(1). However, with faster technologies and significantly lower sequencing costs, new genome analyses are produced(2). Nowadays it takes only a day in a well-equipped laboratory(3).

Figure 1. The picture shows an illustration of CNV on a genome. The right side shows a duplication, left side shows a deletion event.



Copy number variations

Copy number variations fall into the category of structural variants. These are mutations in DNA, which may have beneficial, neutral, or negative effects on the organism. It is a change in an organism's chromosome, involving a DNA fragment that is approximately 1 kb or larger, therefore we consider it a larger mutation event(4). Even though they can be beneficial in the process of evolution, they can cause various syndromes and disorders. Variants are further divided into

translocations, insertions, inversions and finally CNVs. These variants are visualized on **Figure 1**. (CNVs) and **Figure 2**. (translocation, insertion, inversion). Whereas CNVs change the amount of a certain genomic range, it can be detected via low-coverage sequencing, mapping, and read counting. On the other hand, translocations and inversions do not add or remove any genomic material (only shuffle it), thus to detect them, we need to find the precise breaking points. This is possible only with a deep-coverage (>30x) sequencing and even then it does not have great accuracy[5,6]. Insertions bring new genomic material (usually small), which does not have a copy previously on the genome, thus the breaking points and the whole inserted sequence need to be found by deep sequencing. Thus, this article will focus only on CNVs and on methods for their detection.

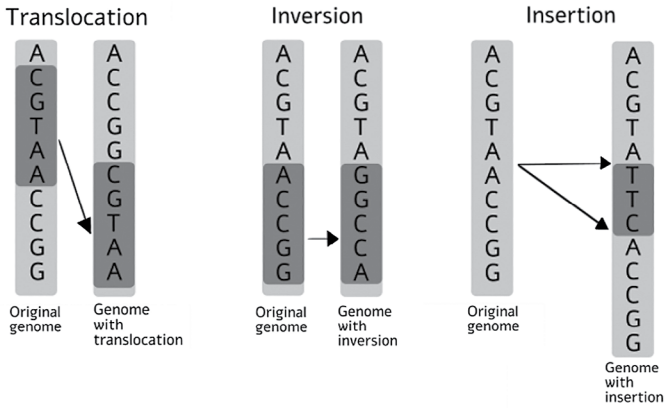
CNV or Copy Number Variant is a phenomenon that represents a significant source of genetic diversity among different species including humans. For instance, it was discovered Higher *AMY1* copy numbers improve the digestion of starchy foods and as a result this gene is related with a history of diet-related selection pressures[7]. However, CNVs are associated with various syndromes and diseases as well[8]. CNVs, particularly common and frequently found in healthy populations, contribute to the inception of cancer as well. For instance, researchers have found that a small deletion in *Mtus1* gene is associated with a decreased risk of familial breast cancer[9]. Another cancer-associated CNV was found in the gene *MLLT4* which seems to be linked with the Li-Fraumeni cancer predisposition disorder[10]. They are associated with schizophrenia, autism, or susceptibility to HIV infection[11]. Moreover, CNVs can cover part of a gene, whole gene, or even several genes and therefore they are likely to have a role in the alternation of human physiological functions, which are essential processes such as metabolism, movements, reproduction, etc.[8,12]. There have been various methods developed with an aim to detect these mutations and even successful pioneer attempts to remedy them, for example the medication Zolgensma[13] which reverts the CNV loss of gene SMN1 or its dysfunction by inserting a new working copy to the genome of an infant. Although this medication is still in its early stages, does not work perfectly, and costs $2.1million for a single treatment, it paves the way for the targeted one-time gene therapy in the field of personalized medicine.

Methods for the detection of CNVs

First generation of sequencing

Over the years many different methods have been developed for CNV detection. The first methods were cytogenetic techniques such as comparative genomic hybridization (CGH), fluorescence in situ hybridization (FISH) and others, which sometimes required visual inspection of chromosomes. Only anomalies of whole chromosomes or variations expanding over a few Mb in size were detected visually, therefore there was a need for a method that would provide a higher resolution than cytogenetics[14]. However, it is important to mention that the development of cytogenetic techniques is a continuous process. For instance, copy number changes have been detected that are as low as 1 Kb[15].

Figure 2. *The image displays illustrations (from left to right) of translocation, inversion and insertion. Left part displays an original genome, the right part shows the same genome with an illustrated structural variant.*



Smaller mid-sized CNVs between 5–500 kb could be detected by a molecular biology method called Southern-blotting[14]. However, this approach proved to be a laborious and time-consuming method that requires large amounts of high-quality DNA. Finally, with the arrival of amplification-based PCR methods, an analytical resolution to single nucleotides was available[16].

Next-generation-based CNV detection methods

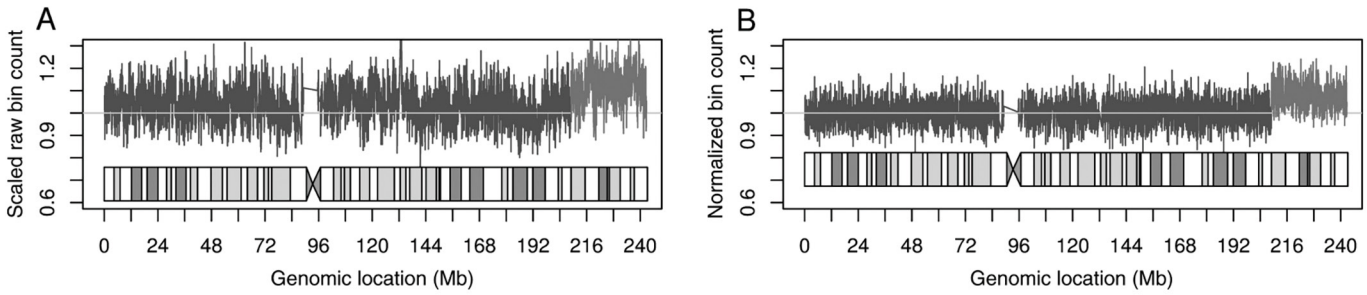
All of the above-mentioned methods were part of the first generation of sequencing and they were considered a golden standard in DNA diagnostics, yet detection of CNVs showed to be challenging with these methods. Fortunately, the next-generation-based CNV analyses managed to overcome some of the mentioned limitations. This new technology brought valuable tools for clinical diagnostics of genomic variations, including CNVs.

Nowadays, next-generation sequencing (NGS) represents a valuable tool for clinical diagnostics and provides a sensitive and accurate approach for the detection of the major types of genomic variations, including CNVs. Three main strategies for CNV analyses based on NGS technologies are whole-genome, whole-exome, and targeted sequencing. Whole-genome sequencing analyses the genome, whereas whole-exome only exomes, which take up roughly one percent of the genome[17]. Targeted sequencing, as the name suggests, analyses only selected sets of genes or genomic regions. Due to the smaller sequenced part of the genome, targeted and whole-exome sequencing usually provide deep-coverage data, whereas whole-genome approaches often rely on shallow sequencing to keep the processing costs low. Thus, the tools for structural variation detection greatly differ for these two kinds of strategies (deep-coverage, small part of genome vs shallow-coverage, whole-genome).

Detection of CNVs using shallow whole-genome sequencing

Whole-genome sequencing (WGS) is the analysis of the entire genomic DNA sequence of an organism at a single time. WGS analysis runs under the support of next-generation sequencing technologies. This technology reduced the

Figure 3. *Bin counts of a particular chromosome before and after normalization. The x-axis shows the coordinate of each bin in Mb. The y-axis in picture A is scaled bin count and in picture B it stands for normalized bin count. The two lighter sections in both pictures display a place where CNV occurred. As can be seen, after normalization the mentioned region is more accentuated.*



expenses for the WGS method and thus it became a comprehensive method for analyzing entire genomes. Shallow whole-genome sequencing is a specific method that obtains genomic data using low-coverage most frequently between 0.1x and 0.5x (coverage of 1x means that every base is read once in average). It is an even more cost-effective approach, it provides more data, greater statistical power and new rare variant discovery capabilities, while its accuracy is very reasonable[18]. CNV is a type of structural variant that can be detected using this method of sequencing even though the coverage can be quite low.

Overall, detection of any CNV is limited by these factors[19]: the size of the particular CNV, coverage, biological and technical variability of the event region. Naturally, it depends on the type of analysis that is being performed. Some detections are done on more challenging data, for example non-invasive prenatal testing, where other factors such as the percentage of fetal fraction plays a crucial role in the successful detection of a CNV.

Results from data with higher coverage are more precise and sensitive, however, lower coverage is overall cheaper and faster. In contrast to the length of the CNV, this factor can be directly changed to obtain more accurate results, but with higher production costs. Yet, shallow WGS works with very small coverages and still manages to obtain reliable data[16,20]. Because of this, it proposes various clinical applications from the detection of CNVs that cause development issues to variations causing malignity and mental retardation[21,22].

Biological and technical variability of the event region refers to the fact that some sectors can be more variable than others. It can be caused by various factors such as repetitive elements, mapping ability and so on. As a result, these regions are harder to detect and are usually filtered out from the analyses.

General approach for the detection of CNVs

Naturally, various CNVs detection tools have been developed over the years[23-26]. Each has its own advantages such as fast computation time, accuracy, lucid results, or easy use. However, this depends on the data that they are given, for instance some are designed for low-coverage data and others require at least certain coverage to work properly. However, most of them share some similarities in their approaches. Usually, they require mapped reads to begin with and then they follow four main steps. The mapped reads are separated into smaller sections called bins. Subsequently, normalization and noise correction techniques are applied and finally, the normalized signal is segmented and scanned for CNVs. The whole process is often referred to as CNV-calling.

Binning is a process where a CNV detection tool partitions reads into bins according to their mapped position. Bin-size is the size of a bin or the number of bases inside the bin. Although this parameter can be often adjusted by the user, some tools propose a method to determine the optimal bin-size. The final resolution strongly depends on the bin-size. Larger bin-size results in worse resolution and faster computational time, however, the sensitivity for smaller aberrations decreases. Term bin-count is the number of reads that belong to a particular bin. This number varies between different genomic regions due to different mappability and biological reasons. Thus, it needs to be normalized to obtain bin-count purified from these biological biases.

One of the most important steps is normalization. In this process, the program adjusts measured values to a hypothetical common scale. Normalization reduces noise and biases commonly seen in samples and therefore can considerably change final sensitivity and specificity. The drawback of this step is that a normalization usually needs a lot of genetically healthy samples for training. The effect of normalization can be seen in **Figure 3**[19].

Another major step in CNV prediction is segmentation. The aim is to gather read depth signals with similar intensity and thus "smoothing" out the noisy bin-count signal into levels. Circular binary segmentation (CBS) is a popular method used for segmentation and the majority of the CNV detection tools use it.

Following this, the segments that deviate from the average read-depth signal are considered to be variant regions. However, read depth signals are noisy due to different aspects such as different mapping abilities of the tested sample and reference genome. As a result, variant regions may be falsely identified. The crucial process in this step is distinguishing the spurious variants from the true copy variants or assigning some confidence to the called CNVs. High-confidence CNVs are usually long, have an expected level of signal intensity for a single gain or loss, and are not in the badly mappable regions of the genome.

Biases arising from WGS detection methods

Many systematic biases arise from whole-genome next-generation sequencing data, which are susceptible to create noise in sequencing data. The most commonly seen bias that arises from sequencing is called GC bias[27]. This bias is caused by proportionally different sequencing of regions with different proportions of guanine and cytosine bases (GC content) across the genome; it is one of the known platform's technical limitations. The presence of regions with poor or rich GC content leads to uneven coverage of reads across the genome. Local regression or local polynomial regression (LOESS) is commonly used to deal with GC bias[28]. LOESS regression merges together bins with similar GC content in a certain interval. This correction is applied to every bin-count, or to individual bins as their weights. The GC bias is different for each sequenced sample and thus needs to be corrected within the sample.

The mappability bias is another systematic bias that affects the results of any CNV detection tool. Mappability of a region is the chance that read will be sequenced and mapped successfully to the region. However, particular regions are very hard to sequence and some regions (e.g. repeating regions) are challenging to map thus these regions will have a very low mappability. Regions with mappability under a certain threshold are usually excluded from the CNV prediction. Bins within regions with better mappability are then normalized to the same level of mappability to correct for the mappability bias within them (**see Figure 3**).

The DNA structure can differ between populations and subpopulations. This relation is called population stratification and it is a source of another bias that can affect the detection. Principal component analysis (PCA) normalization is sometimes used to remove this kind of bias. The aim of this method is to reduce the noise commonly seen across populations while still preserving information directly linked to the sample[19,29]. The disadvantage is the need for a high amount of healthy samples for training. Moreover, the used laboratory protocol affects the parameters of noise, thus it usually needs to be retrained for other laboratory protocols and/or sample tissues.

It is important to mention another aspect of CNV detection and that is linked to the significance of individual variations. The controversy revolves around the fact that many microdeletions and microduplications are either hard to detect with satisfiable accuracy or that their impact on the fetus is not yet fully known. Therefore women who choose NIPT, especially for microdeletion/microduplication detection, should be well informed about the accuracy, reliability, false positive and false negative rates, and the limited ability to predict future intellectual development[30].

Conclusion

The development of new methods for CNV detection is continuous and ongoing. The importance of CNV detection is growing since they are essential to better understand the plasticity of our genome and to elucidate its possible connections to various diseases. The smaller CNVs are often innocuous, however larger ones (500 kb and more) can be linked to the development of disorders and cancer[31]. Methods that have been developed helped to reveal these consequences and enabled us to prepare for the right treatment. Moreover, technologies of next-generation sequencing sped up the detection process due to their ability to compute large amounts of data. Thus the whole-genome sequencing could be done more effectively, which led to new methods and approaches. New convenient tools are being designed for storing, searching, annotating, and evaluating CNV-related data. Resolution of such tools is increasing and subsequently, new variants are discovered. This rapid development leaves a challenge for numerous genetic institutions, researchers, laboratory diagnosticians, and other related corporations in the form of correct interpretation of CNVs evolutionary significance and especially their clinical impact on humans.

Acknowledgment

This article was created with the support of the OP Integrated Infrastructure for the projects: ITMS: 313011V578 and 313011V446, both co-financed by the European Regional Development Fund.

Poďakovanie

Tento článok vznikol za pomoci OP Integrovanej Infraštruktúry pre projekty: ITMS: 313011V578 a 313011V446, spolufinancované zo zdrojov Európskeho fondu regionálneho rozvoja.

REFERENCES

1. Venter JC, Smith HO, Adams MD. The Sequence of the Human Genome. Clin Chem. 2015;61: 1207–1208.
2. Consortium IHGS, International Human Genome Sequencing Consortium. Correction: Initial sequencing and analysis of the human genome. Nature. 2001. pp. 565–566. doi:10.1038/35087627
3. Samuelsson T. The Human Genome in Health and Disease: A Story of Four Letters. Garland Science; 2019.
4. Bruno A, Aury J-M, Engelen S. BoardION: real-time monitoring of Oxford Nanopore Technologies devices. doi:10.1101/2020.06.09.142273
5. Park D, Park S-H, Ban YW, Kim YS, Park K-C, Kim N-S, et al. A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnol. 2017;17: 1–8.
6. Dong Z, Jiang L, Yang C, Hu H, Wang X, Chen H, et al. A robust approach for blind detection of balanced chromosomal rearrangements with whole-genome low-coverage sequencing. Hum Mutat. 2014;35: 625–636.
7. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39: 1256–1260.
8. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10: 451–481.
9. Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, et al. Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. Carcinogenesis. 2007;28: 1442–1445.
10. Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, et al. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proc Natl Acad Sci U S A. 2008;105: 11264–11269.
11. Liu S, Yao L, Ding D, Zhu H. CCL3L1 Copy Number Variation and Susceptibility to HIV-1 Infection: A Meta-Analysis. PLoS One. 2010;5. doi:10.1371/journal.pone.0015778

12. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Medicine. 2018. doi:10.1186/s13073-018-0606-6
13. Zolgensma - one-time gene therapy for spinal muscular atrophy. Med Lett Drugs Ther. 2019;61: 113–114.
14. Pös O, Radvanszky J, Styk J, Pös Z, Buglyó G, Kajsik M, et al. Copy Number Variation: Methods and Clinical Applications. NATO Adv Sci Inst Ser E Appl Sci. 2021;11: 819.
15. Evangelidou P, Alexandrou A, Moutafi M, Ioannides M, Antoniou P, Koumbaris G, et al. Implementation of high resolution whole genome array CGH in the prenatal clinical setting: advantages, challenges, and review of the literature. Biomed Res Int. 2013;2013: 346762.
16. Kucharik M, Gnip A, Hyblova M, Budis J, Strieskova L, Harsanyova M, et al. Non-invasive prenatal testing (NIPT) by low coverage genomic sequencing: Detection limits of screened chromosomal microdeletions. PLoS One. 2020;15: e0238245.
17. Website. Available: https://www.mlo-online.com/molecular/dna-rna/article/13017563/wes-vs-wgs-why-the-exome-isnt-the-whole-story-and-sometimes-when-its-better.
18. Shallow Whole Genome Sequencing. [cited 29 Oct 2021]. Available: https://www.cd-genomics.com/shallow-whole-genome-sequencing.html
19. Zhao C, Tynan J, Ehrich M, Hannum G, McCullough R, Saldivar J-S, et al. Detection of Fetal Subchromosomal Abnormalities by Sequencing Circulating Cell-Free DNA from Maternal Plasma. Clinical Chemistry. 2015. pp. 608–616. doi:10.1373/clinchem.2014.233312
20. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res. 2014;24: 2022–2032.
21. Hehir-Kwa JY, Wieskamp N, Webber C, Pfundt R, Brunner HG, Gilissen C, et al. Accurate distinction of pathogenic from benign CNVs in mental retardation. PLoS Comput Biol. 2010;6: e1000752.
22. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1: 62.
23. Straver R, Sistermans EA, Reinders MJT. Introducing WISECONDOR for noninvasive prenatal diagnostics. Expert Rev Mol Diagn. 2014;14: 513–515.
24. Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res. 2019;47: 1605–1614.
25. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12: e1004873.
26. Dharanipragada P, Vogeti S, Parekh N. iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. PLoS One. 2018;13: e0195334.
27. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE. 2013. p. e62856. doi:10.1371/journal.pone.0062856
28. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics. 2009. pp. 1061–1067. doi:10.1038/ng.437
29. Principal Component Analysis and Factor Analysis. Principal Component Analysis. pp. 150–166. doi:10.1007/0-387-22440-8_7
30. Rose NC, Benn P, Milunsky A. Current controversies in prenatal diagnosis 1: should NIPT routinely include microdeletions/microduplications? Prenat Diagn. 2016;36: 10–14.
31. Valsesia A, Mace A, Jacquemont S, Beckmann JS, Kutalik Z. The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. Front Genet. 2013;0. doi:10.3389/fgene.2013.00092

Zuzana Klinovská
Za Valy 473/2
957 01 Bánovce nad Bebravou
e-mail: z.klinovska@gmail.com