Comparative Exome Capture Methods to Investigate Genes Involved in Hypopituitarism in a Brazilian Population

Anna Flavia Figueredo Benedetti (anna.benedetti@usp.br)
Laboratorio de Hormonios de Genetica/LIM42, Hospital das Clinicas da Faculdade de Medicina da Universidade de São Paulo
Qiany Ma
Department of Human Genetics, University of Michigan
Jun Li
Department of Human Genetics, University of Michigan
Ayse Bilge Ozel
Department of Human Genetics, University of Michigan
Sally Camper
Department of Human Genetics, University of Michigan
Berenice Bilharinho Mendonca
Laboratorio de Hormonios de Genetica/LIM42, Hospital das Clinicas da Faculdade de Medicina da Universidade de São Paulo
Antonio Marcondes Lerario
Department of Internal Medicine, Division of Endocrinology, Metabolism and Diabetes, University of Michigan
Ana Carolina Tahira
Instituto Butantan
Luciani Renata Carvalho
Laboratorio de Hormonios de Genetica/LIM42, Hospital das Clinicas da Faculdade de Medicina da Universidade de São Paulo

Research Article

Keywords: Whole Exome Sequencing, hypopituitarism, molecular diagnosis

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-244377/v1

Abstract

Whole Exome Sequencing (WES) has been a useful tool to improve molecular diagnosis in hypopituitarism, leading to the discovery of at least 8 new genes in the last 7 years. However, some genes associated with hypopituitarism show low coverage in this methodology, limiting its use for molecular diagnosis. Our objective is to compare three library prepping kits, NimbleGen (Roche), SureSelect (Agilent) and Nextera (Illumina) examining the best performance related to sequencing quality, exon extension coverage (≥98%) and base depth read (≥20x) of 44 genes associated with hypopituitarism and 32 involved in pituitary development. Three different groups composed of 2 HapMap samples (Group 1), 2 Brazilian patients with hypopituitarism and their respective mothers (Group 2) and 109 random Brazilian samples (Group 3) were sequenced in Illumina platform. Group 1 and 3 were performed using all three library prepping kits, while group 2 was performed with NimbleGen and SureSelect. Although all technologies covered the selected genes with similar efficiency regarding poor (less than 20%) and rich (more than 80%) GC areas, SureSelect has shown to reach the most uniform coverage in the selected region with a lower level of duplicate reads, as well as a higher number of identified pathogenic variants.

Introduction

Combined Pituitary Hormone Deficiency (CPHD) is the deficiency of one or more pituitary hormones, affecting 1:8000 births worldwide[1]. It may lead to short stature, weight gain, infertility, among other problems depending on which hormones are not being produced. It can be either idiopathic or congenital, and either non-syndromic, leading only to pituitary hormonal deficiencies, or syndromic, with extra-pituitary phenotypes, such as septo-optic dysplasia and holoprosencephaly[2].

Several genes have been reported carrying mutations leading to CPHD and most of them were described by using the gene candidate approach by using the Sanger Method. However, all the genes described so far explain only a small percentage (around 15%) of the patients' clinical features, as both Fang et al[2] and DeRienzo et al[3] have pointed out. Nowadays, the method of choice is Whole Exome Sequencing (WES), capable of sequencing every coding region of the genome, allows for the researcher to investigate known genes as well as finding novel variants in yet unknown ones, increasing the possibility of reaching a diagnosis for these patients.

We used the NimbleGen kit Ez v3 to prepare DNA samples for WES of 23 patients with idiopathic hypopituitarism (11 isolated cases, 12 families). However, it was noted that some genes had a lower coverage than what was expected for a trustworthy variant calling, making it impossible to analyze these regions. The lack of proper coverage in these genes may be due to high GC content, as this is a known source of bias originating from the necessary PCR step in library preparation[4]. This uneven coverage can also be observed in databases such as gnomAD, which has graphs showing median coverage of genes in big sample groups.

Previous studies have already investigated different methodologies involving prepping kits[5–8]. Therefore, to elucidate which technology better covers important regions for proper molecular diagnosis of CPHD patients, we decided to compare different library prepping kits for WES for a set of known genes. This region is comprised by those already identified in published studies as causing CPHD, along with 32 genes with no known pathogenic mutations related to hormone deficiency, but having a role in pituitary development during embryogenesis[9].

Results

Sequencing Quality, Duplicate reads, Probe Analysis of genes involved in CPDH or pituitary development, GC regions, Coverage, Variant Calling and Public databases will be presented as the results in three studied groups (Group 1, comprised of 2 Japanese HapMap samples sequenced by Shigemizu *et al*[5]; Group 2, comprised of 2 patients with hypopituitarism and their mothers; and Group 3, of random 109 Brazilian samples (Figure 1).

Sequencing Quality

Sequencing quality was analyzed to check whether all samples had comparable Phred scores and number of sequenced reads. Although FASTQC[10] showed that sequencing quality was proper and medium reads for Group 1 were similar to each other, Group 2

had a great difference in the number of sequenced reads, with SureSelect attaining 97 million reads and NimbleGen only 69 million. To compare raw base depth considering only these parameters, SureSelect's was 93.95x and NimbleGen's 54.25x**.** Group 3 showed large variation among its sequenced reads, SureSelect with 89 million reads (raw base depth of 99.39x), Nextera with 91 million (95.18x) and NimbleGen 70 million (55.75x) (Table 1).

Despite the divergent raw coverage among the technologies used, it is possible to observe that even at low coverages of 1x and 5x, SureSelect can capture a higher amount of intended target in all 3 groups, and at higher depths, this tendency was clearer. A fairer comparison can be observed in Group 1, which had the least variation in raw reads sequenced and mean coverages >100x in all three technologies, where at higher depths SureSelect (90.02% at 50x) can cover more percentage of bases than Nextera (69.77% at 50x) or NimbleGen's (86.25% at 50x). This data suggests that SureSelect approach can capture target in a more homogeneous aspect than the other two technologies.

Duplicates

Estimating the number of PCR duplications among technologies is an important step to check the bias in target covered regions. PCR duplicated reads could lead to error base-calling variants, being a byproduct of the PCR step, which is applied in library construction in all technologies studied here. The use of this parameter to evaluate these technologies could indicate a good horizontal coverage of regions with a low cost per sequencing. NimbleGen's duplication was the biggest in CPHD genes than any other kit, while SureSelect's duplication was the lowest, although it showed larger variance of duplication rate among samples (Figure 2). This result reinforces that SureSelect approach showed less uneven coverage than NimbleGen or Nextera.

Probe Analysis of 76 genes involved in CPDH or pituitary development during embryogenesis

It is important to check whether all the used methodologies had designed probes to our region of interest, which span across 161,022 bp, as this may result in better coverage in some genes. However, it was shown that although not all regions had probes specifically designed for them, the entire region was covered by nearby probes. The overlapping probes in these regions can be seen in Figure 3.

GC regions

Investigation of the coverage in regions regarding GC content show that all methodologies have bias regarding GC areas, rich (>80%) or low (<20%). Outside CG-rich regions, NimbleGen shows less depth than the other methodologies, which could be an effect of lower mean coverage out of overall coverage (Figure 4). However, it is important to observe that Nextera showed a preference in covering lower over higher GC-rich regions. In our context this is important, because comparing the 76 genes of interest in this study to 76 random genes from the genome, it is possible to see that our chosen group does have a bigger frequency of higher GC areas. However, it is not statistically different from the random gene group (Figure 5).

Coverage

Regarding overall coverage, the best out of the three kits was Agilent's SureSelect, which had good coverage both for the entire exonic region as well as our region of interest, as seen in graphics on Figures 6 and 7. Here, the graphics show that at 20x, which is the reliable depth for variant calling, SureSelect shows the highest percentage coverage across all comparisons, most importantly this is also observed in our region of interest that targeted 76 genes. While Illumina's Nextera maintained a lower coverage in both gene groups, Roche's NimbleGen had a slight fall in coverage for our regions of interest. However, in regard to the whole WES region, it was comparable to SureSelect.

Variant Calling

The main goal of a researcher when using WES is to find variants that can explain the patient's phenotype. Usually, the focus lies on exonic or splice site regions, as they have a higher probability of having impact on the resulting protein and thus being deleterious.

All technologies are rather similar in the number of called variants in all regions, both whole exome region and specific CPHD genes, although Nextera seems to have a higher number of called variants in out-of-exon regions, such as intronic, downstream, and upstream (Table 2; Table 3).

Public databases

ClinVar is a public archive of the relationship between human phenotypes and genomic variations with supporting evidence, facilitating the association between human variation and clinical findings[11]. For such, when submitting new evidence, users must include the clinical significance according to ACMG criteria[12]. We determine whether these technologies can cover every known pathogenic variant in hypopituitarism genes, so as not to miss any probable cause of the studied phenotype.

ClinVar presents 1808 pathogenic or likely pathogenic variants in the 76 genes here studied. Mean coverage of each loci was performed to check which sequencing kit was able to cover the most of these variants at least 20x. In all the 3 groups, the SureSelect library has a lower number of uncovered variants (22 variants out of 1808). Similarly, NimbleGen library had 80 not covered in any of the sample groups. A quick summary of this information can be seen in Table 4, and for more detail on these variants and their loci, they can be found in Supplementary Information Table S1 of this paper.

ABraOM is a variant repository with the frequency of variants found in a normal Brazilian population. Currently, it consists of Whole Genome Sequencing of 1,171 unrelated elderly individuals [13]. In an earlier version, composed of WES of 609 elderly individuals; 207,621 variants appeared only in this repository, which are then believed to be exclusive to the Brazilian population[14].

Since our goal involves the efficiency of different sequencing library preparation kits in a Brazilian population specifically, we analyzed the variants found in ABraOM patients in the 76 genes studied, out of which 175 were exonic and found to be exclusive of the Brazilian population (Figure 8). Across all 3 groups, SureSelect was the library with lowest number of uncovered variants (Table 5); Supplementary Information Table S2.

Discussion

When using high throughput sequencing technologies, it is necessary to perform quality and coverage analysis before variant filtering, to ensure reliable results. Generally, the coverage of genes known to cause the phenotype is not discussed in published articles that report new variant findings in hypopituitarism. This fact, along with experiences with low coverage in WES sequencing in some of our samples using the NimbleGen kit, which had to be remade, led us to compare the efficiency of other kits available to the general market and to us. As many other comparisons on these kits have already been made[5–7], we decided to focus our comparison in important regions to the disease we have been studying, as to shed light to researchers in this field which approach is better to use in their cohort. For that, we selected genes that are important to pituitary development during embryogenesis, as well as genes that have been associated with hypopituitarism[15–30].

The use of simpler technologies such as gene panels, can be tempting regarding tricky parts of the genome such as the one mentioned in this study, but a low number of molecular diagnosis has been reached according to the literature. Nakaguma *et al.* had a 4% success rate in diagnosing 117 patients using a custom gene panel with 26 genes previously related to hypopituitarism[31]. However, as stated by the author, this was a cohort previously screened and the use of gene panels may return a higher success rate (closer to 15%) if used in a cohort naïve of diagnostic approach, similar to the number found in the overall diagnostic rate for CPHD patients[3,31]. Even so, the approach of using WES is perhaps a better option, since it gives way to the discovery of new genetic causes[2,3].

We also opted to broaden our samples groups and, unlike other comparisons made previously, added different group samples, such as a patient with the disease in question and random Brazilian samples. This was done to ensure that different known biases common to the technique of WES, such as sample, run or laboratory bias were not a big factor on the obtained results[4]. Unlike other populations, few information about the Brazilian population is available in the literature, as evidenced by the only two existing databases containing samples from this group, ABraOM and SELA[14,32].

As all technologies studied are of great quality and achieve their goals, the answer to the question of which is best and should be used comes down to specific parameters and depends on the researcher's targets[7]. For most investigators of WES in regards to medical sciences, the small difference in coverage of coding regions is of great importance, as it directly reflects the ability to identify rare variants[5]. This is also the case for most researchers trying to obtain molecular diagnosis for CPHD patients.

Our results come in contrast to the findings of Clark *et al*, that report that the densely packed and overlapping baits of Roche's NimbleGen granted a higher coverage of targeted regions with a slightly higher edge in sensitivity for SNPs and indels[7]. However, it should be noted the use of different versions of library preparation kits, as theirs was v2.0 of the kit while ours was v3.0, which may explain this difference. This is further exemplified by Asan *et al*, who concluded that between NimbleGen v1.0 and SureSelect All Human Exon, that the latter had a higher number of SNPs[8]. Both studies noted that NimbleGen needed a lower number of reads to reach the expected coverage, which is corroborated by our results, as it reached comparable coverage to the other kits even with a lower number of sequenced reads[7,8].

It was also noted by other studies that Illumina's Nextera had an increase in read depth in areas with 40 to 60% of GC content[5,6]. This, however, did not translate to a higher coverage in genes implicated in CPHD with a high GC content, such as *SOX3*. In fact, it presented with the lowest coverage among the kits. This may be due to its fragmentation being done by enzymes, which has a greater fragment bias since it is not random shearing like in mechanical fragmentation. Therefore, other kits that use mechanical shearing for library preparation may have a more adequate coverage in these regions.

Lastly, Agilent's SureSelect All Human Exon v5's higher coverage in coding regions is seen across different comparison studies, as well as here[5,6]. Not only it reached a higher expected coverage in the whole exonic region, but also for our specific region of hypopituitarism genes and those present in pituitary development. We compared coverage of known pathogenic or likely pathogenic *loci* in our region of interest found on ClinVar across kits, as well as of *loci* related to Brazilian polymorphisms according to ABraOM. In both cases, SureSelect had the best number of *loci* covered. Therefore, it is a strong contender for the best kit out of the three.

In conclusion, when comparing library preparation kits for WES taking into consideration studies looking for molecular diagnosis of CPHD patients, Agilent's SureSelect kit has the best performance. Moreover, regardless of the methodology used, it is of utmost importance to properly analyze whether every known causative gene has been properly covered in the sequenced samples, so as not to miss variants.

Methods

Editorial Policies and Ethical Considerations

This study was approved by CEP (Comitê de Ética em Pesquisa) and CONEP (Comissão Nacional de Ética em Pesquisa) ethics committees under the number CAAE 06425812.4.0000.0068. All participants or their guardians signed a written form of informed consent agreeing to be involved in the study, according to resolution CNS 466/12.

Samples and sequencing

First, we analyzed two HapMap samples, both prepared with each of the technologies Nextera (Illumina Inc.), SureSelect v5 (Agilent Inc. Santa Clara, CA, USA) and NimbleGen SeqCap EZ v3 (IntegenX Inc., Pleasanton, CA, USA), that have been sequenced and made available by Shigemizu *et al.* at the DNA Data Bank of Japan (DDBJ) under DRA003736[5]. Each of the technologies differ in size of target region, spanning from 45 Mb (Nextera) to 50 Mb (SureSelect) and 64 Mb (NimbleGen). Secondly, four Brazilian individuals (patients I and II and their mothers) were sequenced using NimbleGen and SureSelect v5 technologies for comparison between both technologies. The corresponding data was available at SRA under PRJNA686987. Lastly, a randomized cohort of Brazilian patients were prepared with Nextera (20 samples), NimbleGen (43) and SureSelect v5 (21) following kit's protocol and were analyzed in the same way as previous groups. NimbleGen's sequencing was performed at the Sequencing Core of University of Michigan in collaboration with Dr. Sally Camper, while the other kits were sequenced at Hospital das Clínicas

University of São Paulo's SELA (Sequenciamento em Larga Escala), following manufacturer's protocols specific for each kit. A breakdown on the sample groups analyzed can be found in Figure 1.

Chosen region

From searching the literature, we have selected 76 genes, shown in Table 6, that either are present in pituitary embryogenesis or have mutations found in CPHD patients, despite level of evidence when their variants were classified using ACMG criteria, as shown in Table 7. This region is referred to as "our region of interest" in the text. Meanwhile, the whole exome region each kit targets for sequencing may be referred to as "global region".

Bioinformatics Analysis

Sequencing quality was checked using FASTQC software (v.0.11.2)[10], followed by alignment using BWA (v.0.6.1-r104)[33] with hg19 assembly as reference genome. PCR Duplications removal was performed using samtools (v.1.6)[34] or Picard (v.2.18.2)[35] algorithms. Realignment and recalibration were performed using GATK version 2.8, while coverage analysis was performed using the softwares BedTools (version 2.25.0)[36] and Qualimap2 (version 2.2.1) [37]. Finally, variants were called using GATK's HaplotypeCaller v3.2.2[38] and annotated with SnpEff[39].

A comprehensive visualization of gene by gene coverage was performed with R function plot.baseCoverage, which can be found in github (https://github.com/anna-benedetti9/plot-basecoverage). This function takes as input a bed file created with bedtools coverage -d and plots the coverage of any given gene found in the file.

Lastly, the analysis described by Naslavsky *et al*[14] was applied to filter for quality variants in the AbraOM database. Any variant that appeared in any other populational databases such as 1000 Genomes, ESP6500, gnomAD and ExAC was excluded to filter for variants that are exclusive to the normal Brazilian population in our region of interest.

Declarations

Data availability

HapMap samples analyzed as Group1 were made available by Shigemizu *et al*[5] at the DNA Data Bank of Japan (DDBJ), under the number DRA003736. Group 2's Brazilian patients with CPHD and their mothers are available at SRA under PRJNA686987. All other data used in this study is available upon request with the corresponding author.

Acknowledgements

The authors would like to thank Helena Brentani's laboratory support for the computational infrastructure to perform the analyses in the multiuser cluster.

References

1. Medicine®, U. S. N. L. of. Genetics Home Reference. NF2 (2007).
2. Fang, Q. *et al.* Genetics of Combined Pituitary Hormone Deficiency: Roadmap into the Genome Era. *Endocr. Rev.***37**, 636–675 (2016).
3. De Rienzo, F. *et al.* Frequency of genetic defects in combined pituitary hormone deficiency: a systematic review and analysis of a multicentre Italian cohort. *Clin. Endocrinol. (Oxf).***83**, 849–860 (2015).
4. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.***12**, R18 (2011).
5. Shigemizu, D. *et al.* Performance comparison of four commercial human whole-exome capture platforms. *Sci. Rep.***5**, 12742 (2015).
6. Chilamakuri, C. S. *et al.* Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics***15**, 449 (2014).

7. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.***29**, 908–914 (2011).
8. Asan *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.***12**, R95 (2011).
9. Rizzoti, K. Genetic regulation of murine pituitary development. *J. Mol. Endocrinol.***54**, R55–R73 (2015).
10. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
11. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.***46**, (2018).
12. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.***15**, 733–747 (2013).
13. Naslavsky, M. S. *et al.* Whole-genome sequencing of 1 , 171 elderly admixed individuals from the largest Latin American metropolis (São Paulo , Brazil) † Corresponding author s ZIP 05508090 ZIP 05508090 Abstract As whole-genome sequencing (WGS) becomes the gold standard tool . 1–29 (2020).
14. Naslavsky, M. S. *et al.* Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum. Mutat.***38**, 751–763 (2017).
15. Webb, E. A. *et al.* ARNT2 mutation causes hypopituitarism, post-natal microcephaly, visual and renal anomalies. *Brain***136**, 3096–3105 (2013).
16. McCabe, M. J. *et al.* Variations in *PROKR2* , But Not *PROK2* , Are Associated With Hypopituitarism and Septo-optic Dysplasia. *J. Clin. Endocrinol. Metab.***98**, E547–E557 (2013).
17. Giri, D. *et al.* Novel FOXA2 mutation causes Hyperinsulinism, Hypopituitarism with Craniofacial and Endoderm-derived organ abnormalities. *Hum. Mol. Genet.***26**, 4315–4326 (2017).
18. Lal, R. A. *et al.* A Case Report of Hypoglycemia and Hypogammaglobulinemia: DAVID Syndrome in a Patient With a Novel NFKB2 Mutation. *J. Clin. Endocrinol. Metab.***102**, 2127–2130 (2017).
19. Starink, E. *et al.* Genetic analysis of IRF6, a gene involved in craniofacial midline formation, in relation to pituitary and facial morphology of patients with idiopathic growth hormone deficiency. *Pituitary***20**, 499–508 (2017).
20. D Hidalgo-Santos, A. *et al.* A Novel Mutation of MAGEL2 in a Patient with Schaaf-Yang Syndrome and Hypopituitarism. *Int. J. Endocrinol. Metab.***16**, e67329 (2018).
21. Simm, F. *et al.* Identification of SLC20A1 and SLC15A4 among other genes as potential risk factors for combined pituitary hormone deficiency. *Genet. Med.***20**, 728–736 (2018).
22. Pereira Ferreira, N. G. B. *et al.* SAT-LB58 Molecular Investigation of Recessive Inheritance by Exome Sequencing of Patients With Congenital Hypopituitarism. *J. Endocr. Soc.***4**, (2020).
23. Smith, J. D. *et al.* Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontonasal dysostosis. *Am. J. Hum. Genet.***95**, 235–40 (2014).
24. Synofzik, M. *et al.* PNPLA6 mutations cause Boucher-Neuhauser and Gordon Holmes syndromes as part of a broad neurodegenerative spectrum. *Brain***137**, 69–77 (2014).
25. Takagi, M., Narumi, S., Hamada, R., Hasegawa, Y. & Hasegawa, T. A novel KAL1 mutation is associated with combined pituitary hormone deficiency. *Hum. genome Var.***1**, 14011 (2014).
26. Tata, B. *et al.* Haploinsufficiency of Dmxl2, encoding a synaptic protein, causes infertility associated with a loss of GnRH neurons in mouse. *PLoS Biol.***12**, e1001952 (2014).
27. Hufnagel, R. B. *et al.* Neuropathy target esterase impairments cause Oliver-McFarlane and Laurence-Moon syndromes. *J. Med. Genet.***52**, 85–94 (2015).
28. Karaca, E. *et al.* Whole-exome sequencing identifies homozygous GPR161 mutation in a family with pituitary stalk interruption syndrome. *J. Clin. Endocrinol. Metab.***100**, E140-7 (2015).
29. Lucas-Herald, A. K. *et al.* A Case of Functional Growth Hormone Deficiency and Early Growth Retardation in a Child With IFT172 Mutations. *J. Clin. Endocrinol. Metab.***100**, 1221–1224 (2015).
30. Bashamboo, A., Bignon-Topalovic, J., Rouba, H., McElreavey, K. & Brauner, R. A Nonsense Mutation in the Hedgehog Receptor CDON Associated With Pituitary Stalk Interruption Syndrome. *J. Clin. Endocrinol. Metab.***101**, 12–15 (2016).

31. Nakaguma, M. *et al.* Novel pathogenic variants in congenital hypopituitarism Genetic diagnosis of congenital hypopituitarism by a target gene panel: novel pathogenic variants in GLI2, OTX2 and GHRHR. (2019) doi:10.1530/EC-19-0085.
32. Lerario, A. M. *et al.* SELAdb: A database of exonic variants in a Brazilian population referred to a quaternary medical center in São Paulo. *Clinics (Sao Paulo).***75**, e1913 (2020).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics***25**, 1754–1760 (2009).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics***25**, 2078–2079 (2009).
35. Broad Institute. Picard Tools - By Broad Institute. *Github* (2009).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics***26**, 841–842 (2010).
37. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics***32**, btv566 (2015).
38. Garrison, E. & Marth, G. *Haplotype-based variant detection from short-read sequencing*. (2012).
39. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).***6**, 80–92 (2012).

Tables

Table 1. Mean sequencing parameters of each technology for each of the 3 groups.

	Group 1			Group 2		Group 3		
	Nextera	NimbleGen	SureSelect	NimbleGen	SureSelect	Nextera	NimbleGen	SureSelect
Total reads	97,307,852	99,959,956	98,589,533	69,915,783	97,503,227	91,929,866	67,156,725	89,435,182
Mapped reads (%)	95,070,415 (97.7)	98,771,192 (98.8)	97,548,776 (98.9)	69,166,641 (98.93)	97,320,814 (99.84)	91,666,213 (99.78)	66,475,108 (98.99)	89,334,131 (99,9)
Duplicated reads (%)	8,070,034 (8.29)	3,949,071 (3.95)	14,931,303 (15.14)	12,267,790 (17.55)	15,413,175 (15.8)	9,276,850 (10.09)	12,388,245 (18.45)	12,102,113 (13.53)
Mapped to region (%)	260,987 (0.27)	209,992.5 (0.21)	325,906 (0.34)	119,715.5 (0.17)	242,078.5 (0,27)	287,133.5 (0.34)	113,787.5 (0.17)	212,250.5 (0.29)
Duplicated mapped to region (%)	70,561 (15.55)	7,858.5 (3.75)	57,284.5 (17.6)	26,160.5 (22.86)	51,946.5 (18.91)	48,493.5 (17.05)	26,936 (24.01)	48,706.5 (9.22)
Mean coverage	102.2	134.2	158.6	53.6	97.11	95.18	53.08	100.28
Coverage 1x (%)	99.67	99.35	99.92	99.6	99.88	99.1	99.41	99.85
Coverage 5x (%)	98.63	97.99	99.75	98.51	99.77	96.78	98.08	99.60
Coverage 10x (%)	96.96	96.90	99.42	97.2	99.55	93.91	96.67	99.01
Coverage 20x (%)	92.08	95.2	98.44	93.12	98.65	87.97	91.47	96.28
Coverage 30x (%)	85.46	93.19	96.96	82.52	96.72	81.64	80.18	90.81
Coverage 50x (%)	69.78	86.25	92.3	46.1	88	68.29	45.9	73.56

	Group 1			Group 2		Group 3		
	Nextera	NimbleGen	SureSelect	NimbleGen	SureSelect	Nextera	NimbleGen	SureSelect
Missense	24,970	24,413.5	24,191.5	25,047.5	24,510.5	21,213.93	22,300.45	21,492
Synonymous	29,114	28,796.5	28,651	30,024.5	30,152.25	26,236.21	27,083.94	26,784
Frameshift indels	714.5	698.5	673.5	593.75	504	565.07	503.29	430
Stop Gain	183	178.5	162	176.75	162.5	163.85	168	130
Stop Loss	74	62.5	64.5	59.25	58.5	56.42	58.25	57
Exonic region	66,190	65,221.5	63,386.5	71,354.25	69,414.75	55,806.57	59,785.84	56,474.5
Intronic region	138,684	138,459	130,390	517,024	753,994	88,117.64	111,574.8	90,008.5
Downstream region	22,767.5	23,510.5	21,144.5	60,635	82,473.5	15,470.14	20,243,94	15,825
Upstream region	18,085.5	18,315.5	17,211	50,803.25	77,520.75	11,732.07	15,331.52	12,197
Splice region	9,990.5	10,079.5	8,857.5	9,228.75	7,957.5	7,157.07	7,728.67	6,718.5
Total	75,925.5	77,682.5	70,153.5	285,339.8	438,573.2	53,185.29	68,378.19	51,722

Table 2. Breakdown of variants by type that were called in each of the groups in the whole exonic region.

Table 3. Breakdown of variants by type that were called in each of the groups in the genes related to CPHD or present in pituitary embryogenesis.

	Group 1			Group 2		Group 3		
	Nextera	NimbleGen	SureSelect	NimbleGen	SureSelect	Nextera	NimbleGen	SureSelect
Missense	76	69	70	83.25	43.5	50.57	54.77	46.5
Synonymous	161	153.5	65.5	159.5	91.5	105.64	82.87	154.5
Frameshift indels	2	2	2	28.5	0	2.35	3.12	3
Exonic region	246	229	99.5	122.25	119.5	151.07	117.32	209.5
Intronic region	5	5	13.5	8.25	74	11.64	138.19	16
Downstream region	13	18.5	9.5	64.5	26.25	35.14	14	13
Upstream region	4.5	7.5	34	20	43.5	7.42	11.06	14.5
Splice region	10.5	2	5	21.75	5.75	68	0	0
Total	74	71	72.5	78.5	78.25	71.28	72.48	76.5

Table 4. Number of variants found in ClinVar described as pathogenic or likely pathogenic in any of the genes related to hypopituitarism with low coverage (less than 20x) for each of the groups (Group 1: HapMap samples; Group 2: Hypopituitarism patients and their mothers; Group 3: Random Brazilin samples)

	Group 1		Group 2			Group 3		
	Nextera	NimbleGen	SureSelect	NimbleGen	SureSelect	Nextera	NimbleGen	SureSelect
Low covered variants	27	83	16	85	8	90	86	22

Table 5. Number of variants found in ABraOM as exclusive to the Brazilian population in any of the genes related to hypopituitarism with low coverage (less than 20x) for each of the groups (Group 1: HapMap samples; Group 2: Hypopituitarism patients and their mothers; Group 3: Random Brazilin samples)

	Group 1			Group 2		Group 3		
	Nextera	NimbleGen	SureSelect	NimbleGen	SureSelect	Nextera	NimbleGen	SureSelect
Low covered variants	15	18	0	21	0	20	19	1

Table 6. Genes found in the literature and their respective entrezID that are used in this study, divided in two groups of those with pathogenic variants related to hypopituitarism and those that are present in pituitary embryogenesis, but no pathogenic variants described to this day.

Related to hypopituitarism (entrezID)		Present in pituitary embryogenesis (entrezID)	
ARNT2 (9915)	KCNQ1 (3784)	AES (166)	SIX3 (6496)
CDH2 (1000)	LEPR (3953)	BMP2 (650)	SIX6 (4990)
CDON (50937)	LHX3 (8022)	BMP4 (652)	SOX4 (6659)
DMXL2 (23312)	LHX4 (89884)	BMP7 (655)	TBX2 (6909)
FGF8 (2253)	NFKB2 (4791)	BMPR1A (657)	TBX3 (6926)
FGFR1 (2260)	OTX2 (5015)	CHD7 (55636)	TCF4 (6625)
FOXA2 (3170)	PAX6 (5080)	CTNNB1 (1499)	TLE1 (7088)
GH1 (2688)	PITX2 (5308)	FGF10 (2255)	WNT4 (54361)
GHR (2690)	PNPLA6 (10908)	FGF18 (8817)	ZSWIM6 (57688)
GHRH (2691)	POU1F1 (5449)	GATA2 (2624)	
GHRHR (2692)	PROKR2 (128674)	GATA3 (2625)	
GHSR (2693)	PROP1 (5626)	GLI3 (2737)	
GLI2 (2736)	RNPC3 (55599)	HES1 (3280)	
GPR161 (23432)	SHH (6469)	HES5 (388585)	
HDAC6 (10013)	SLC15A4 (121260)	HNRNPU (3192)	
HESX1 (8820)	SLC20A1 (6574)	ISL1 (3670)	
HHIP (64399)	SOX2 (6657)	LHX2 (9355)	
IFT172 (26160)	SOX3 (6658)	NOTCH2 (4853)	
IGSF1 (3547)	TCF7L1 (83439)	POLR3A (11128)	
IRF6 (3664)	TGIF1 (7050)	RAX (30062)	
JAK1 (3716)	WDR11 (55717)	RBM28 (55131)	
KAL1 (3730)	WNT5A (7474)	RBPJ (3516)	

Table 7. Variants found in hypopituitarism patients in different studies and each of their ACMG classification according to Varsome.

Gene	Variant	ACMG CLASSIFICATION	Reference
ARNT2	c.1372_1373dupTC/p.S459Ffs*53	Pathogenic	Webb *et al*, 2013
ZSWIM6	c.3487C>T/p.R1163W	Likely pathogenic	Smith *et al*, 2014
PNPLA6	Many	VUS†	Synofzik *et al*, 2014; Hufnagel *et al*, 2015
HNRNPU	c.1615 –1G>A	Pathogenic	Zhu *et al*, 2015
GPR161	c.47T>A/p.L16Q	VUS	Karaca *et al*, 2015
CDON	c.2764G>T/p.E922X	VUS	Bashamboo *et al*, 2016
CHD7	c.2194C>G/p.P732A	Benign	Gregory *et al*, 2013
IFT172	c.5179T>C/p.C1727R; c.337–2A>C	Likely Pathogenic; Pathogenic	Lucas-Herald *et al*, 2015
DMXL2	c.5824_5838delAGTGATGGCAATGGA / p.D1947Sdel	Likely pathogenic	Tata *et al*, 2014
KAL1	c.1704C>A/p.H568Q	VUS	Takagi *et al*, 2014
KCNQ1	c.347G>T/p.R116L; c.1106C>T/p.P369L	Likely pathogenic; Likely pathogenic	Tommiska *et al*, 2017
IRF6	c.697C>T/p.R233C	VUS	Starink *et al*, 2017
NFKB2	c.2596A>C/p.S866R	VUS	Lal *et al*, 2017
FOXA2	c.505T>C/p.S169P	Likely pathogenic	Giri *et al*, 2017
JAK1	8Mb Deletion of Ch 1p31.1 - 1p31.3	VUS	Thakur *et al*, 2017
LEPR		VUS	
SLC15A4	c.1367C>T/p.P456L; c.250C>T/p.L84F	VUS; VUS	Simm *et al*, 2017
SLC20A1	c.266T>C/p.L89S; c.1561C>T/p.L521F	Likely pathogenic	
MAGEL2‡	c.3019C>T/p.Q1007X	Pathogenic	Hidalgo-Santos *et al*, 2018

† Variant of uncertain significance

‡Not included in this study.

Figures

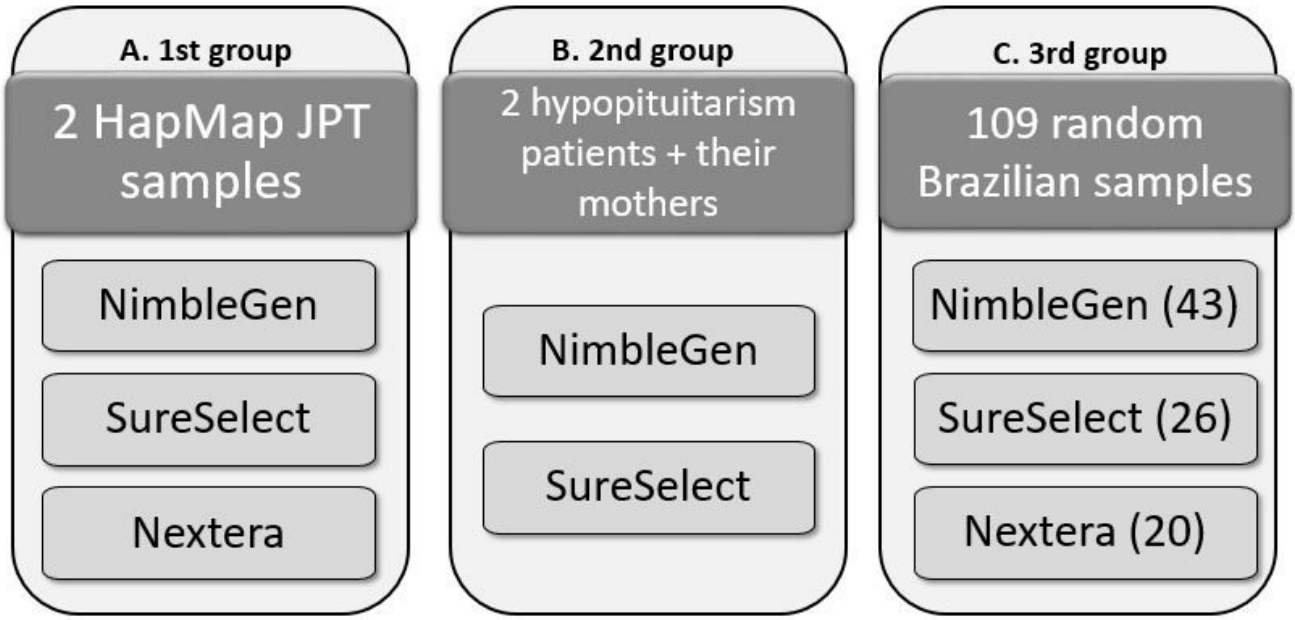

Figure 1

Representation of number of samples prepared with each technology, divided in 3 groups according to the type of samples. (a) Group 1 is comprised of 2 samples that were sequenced using three different technologies; (b) Group 2 is formed by 4 samples sequenced using two different technologies and; (c) Group 3 is comprised of 103 samples that were sequenced by one of the three technologies.

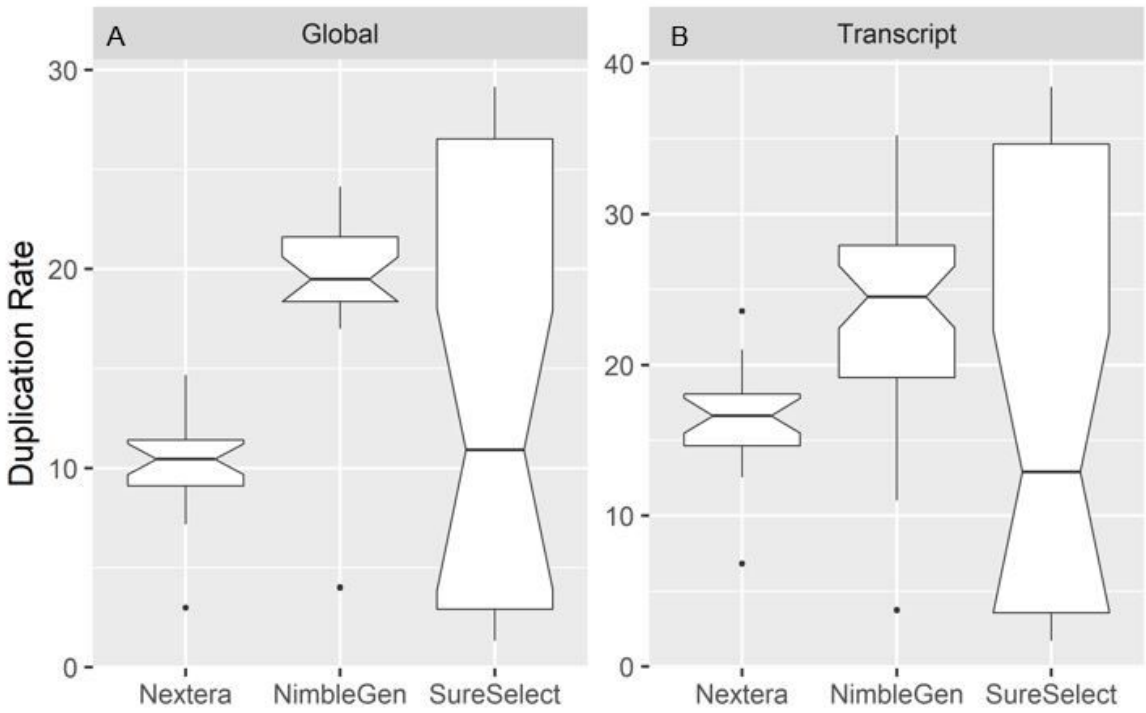

Figure 2

Duplication rates for each of the technologies considering both regions of each technology. The X axis represents each technology, while the Y axis the duplication rate, with representation for the median. Box and whiskers represent interquartile, minimum and maximum values. Line inside the box represent the median value, while outside dots are outliers.

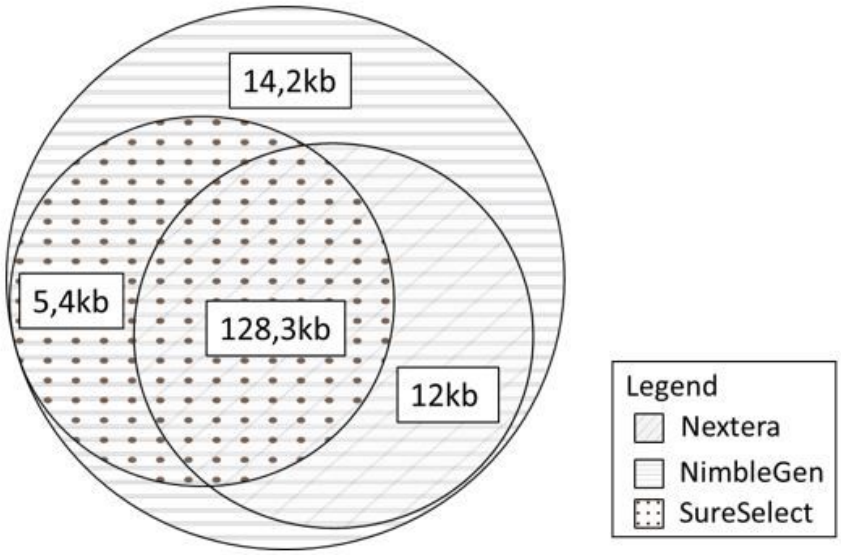

Figure 3

Venn diagram of overlapping probes designed by each technology. Nextera is the only technology with probes designed for the entire region of interest, while NimbleGen has more probes than SureSelect.

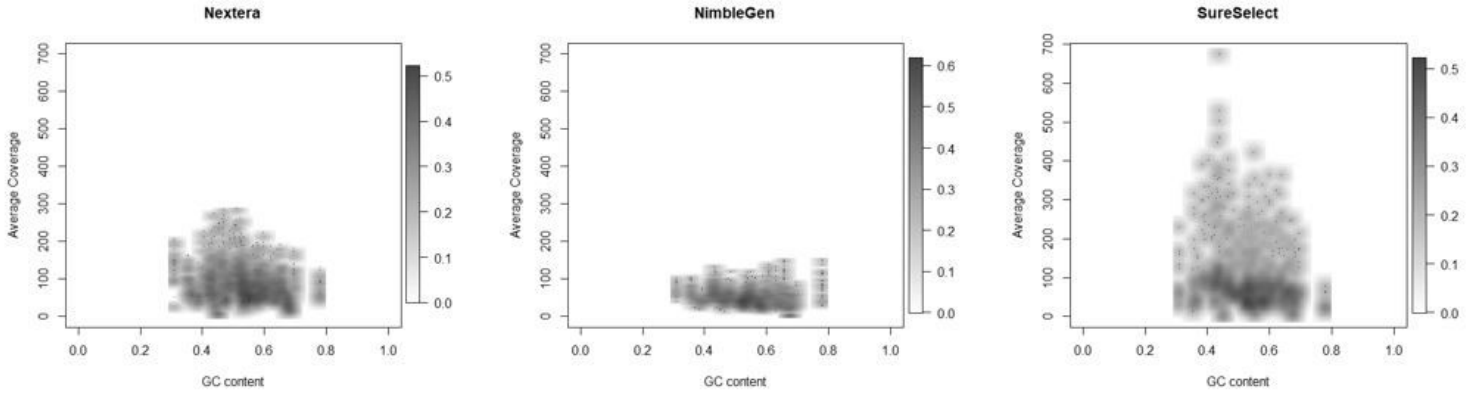

Figure 4

GC content graphs to indicate efficiency in coverage for each of the technologies. The X axis indicates GC content and the Y axis the mean coverage of the kit. The vertical bar shows that the darker the color, higher the number of samples with that coverage. (A) Nextera, (B) NimbleGen and (C) SureSelect.

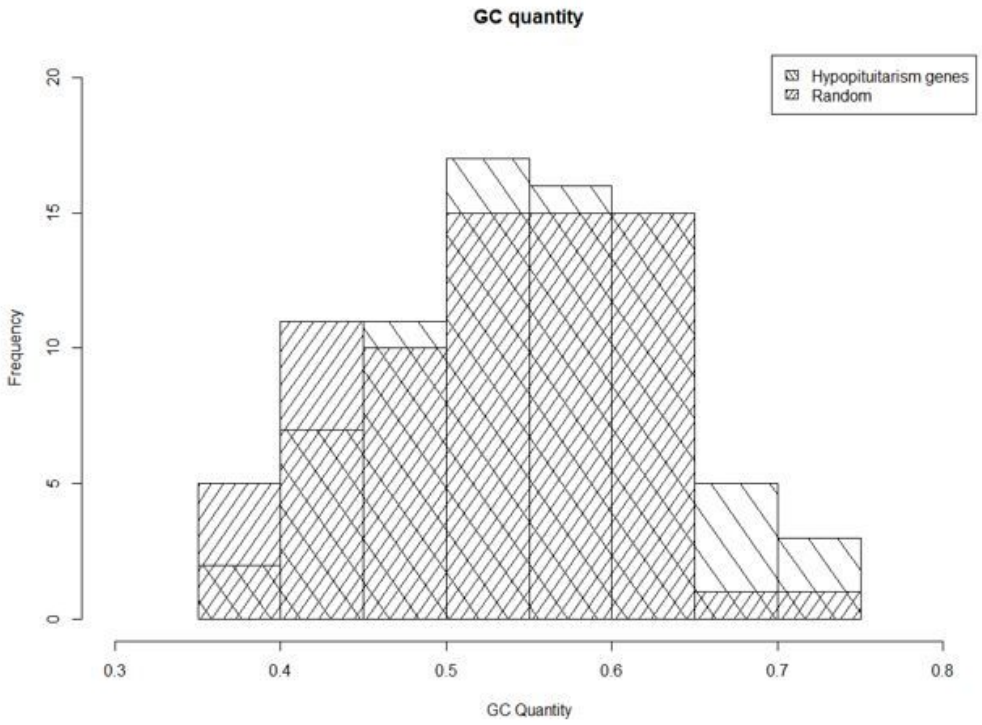

Figure 5

Comparison histogram of GC content in the chosen 76 genes used in the study and 76 random genes selected from the genome. The X axis represents the GC quantity, and the Y axis the frequency of genes.

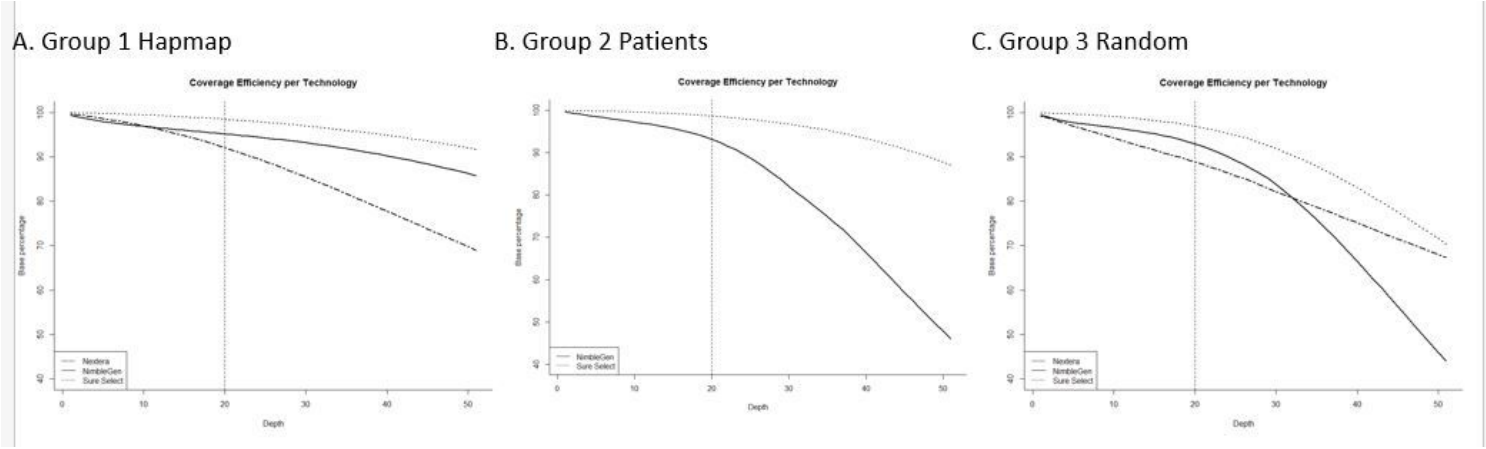

Figure 6

Depth in coverage for the whole exonic region. The X axis represent the depth of coverage, while the Y axis the bases covered in percent. The horizontal line denotes the optimal coverage of 20 times per base. (A) Group 1, (B) Group 2 and (C) Group 3.

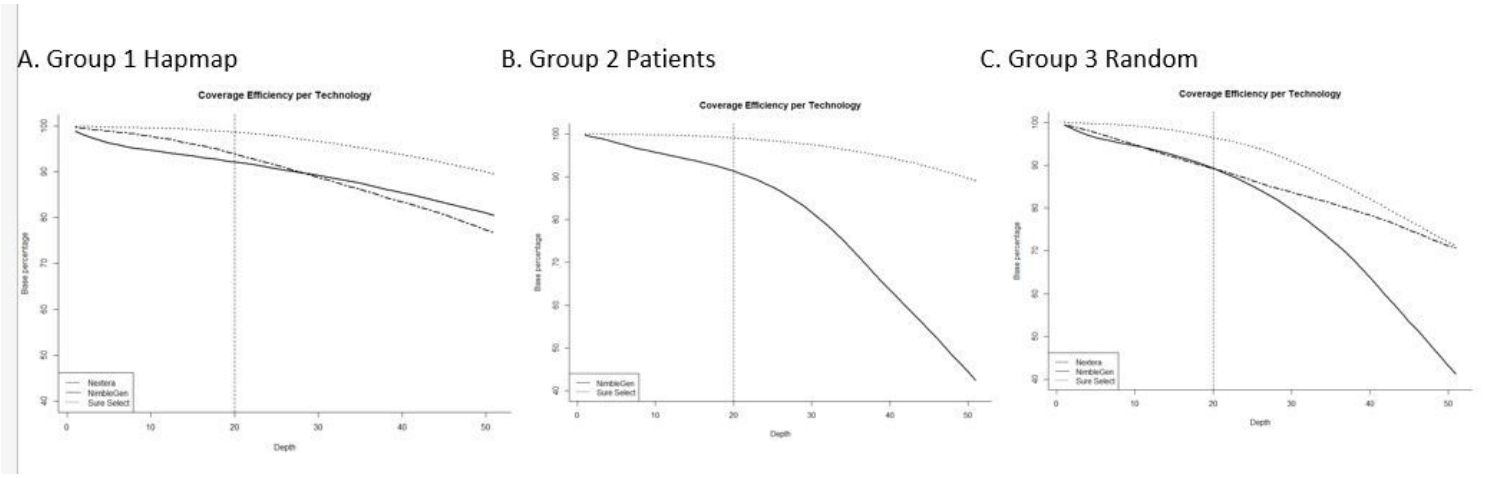

Figure 7

Depth in coverage for the region of interest composed of 76 genes or 161,022 bp. The X axis represent the depth of coverage, while the Y axis the bases covered in percent. The horizontal line denotes the optimal coverage of 20 times per base. (A) Group 1, (B) Group 2 and (C) Group 3.

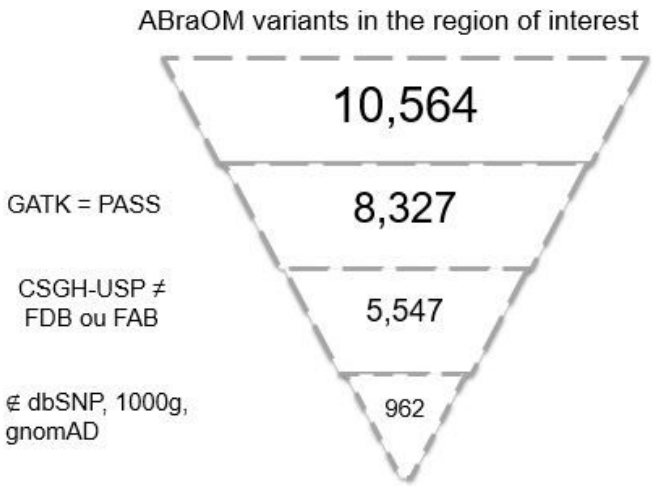

Figure 8

Filtering steps applied to variants found in ABraOM that were located in the region of interest composed of 76 genes. The first two steps are related to base quality (GATK and USP) while the last one is related to frequency of these variants in international databases (dbSNP, 1000 genomes and gnomAD).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx