RESEARCH ARTICLE

Bioinformatic Challenges in Clinical Diagnostic Application of Targeted Next Generation Sequencing: Experience from Pheochromocytoma

Joakim Crona[1]*, Viktor Ljungström[2], Staffan Welin[3], Martin K. Walz[4], Per Hellman[1], Peyman Björklund[1]

1 Department of Surgical Sciences, Uppsala University, SE-75185, Uppsala, Sweden, 2 Department of Immunology, Genetics and Pathology, Uppsala University, SE-75185, Uppsala, Sweden, 3 Departments of Medical Sciences, Uppsala University, SE-75185, Uppsala, Sweden, 4 Department for Surgery and Centre of Minimal Invasive Surgery, Kliniken Essen-Mitte, Academic Teaching Hospital of the University of Duisburg-Essen, DE-45136 Essen, Germany

* joakim.crona@surgsci.uu.se

Abstract

Background

Recent studies have demonstrated equal quality of targeted next generation sequencing (NGS) compared to Sanger Sequencing. Whereas these novel sequencing processes have a validated robust performance, choice of enrichment method and different available bioinformatic software as reliable analysis tool needs to be further investigated in a diagnostic setting.

Methods

DNA from 21 patients with genetic variants in *SDHB*, *VHL*, *EPAS1*, *RET*, (*n*=17) or clinical criteria of NF1 syndrome (*n*=4) were included. Targeted NGS was performed using Truseq custom amplicon enrichment sequenced on an Illumina MiSEQ instrument. Results were analysed in parallel using three different bioinformatics pipelines; (1) Commercially available MiSEQ Reporter, fully automatized and integrated software, (2) CLC Genomics Workbench, graphical interface based software, also commercially available, and ICP (3) an in-house scripted custom bioinformatic tool.

Results

A tenfold read coverage was achieved in between 95-98% of targeted bases. All workflows had alignment of reads to *SDHA* and *NF1* pseudogenes. Compared to Sanger sequencing, variant calling revealed a sensitivity ranging from 83 to 100% and a specificity of 99.9-100%. Only MiSEQ reporter identified all pathogenic variants in both sequencing runs.

Conclusions

We conclude that targeted next generation sequencing have equal quality compared to Sanger sequencing. Enrichment specificity and the bioinformatic performance need to be carefully assessed in a diagnostic setting. As acceptable accuracy was noted for a fully automated bioinformatic workflow, we suggest that processing of NGS data could be performed without expert bioinformatics skills utilizing already existing commercially available bioinformatics tools.

Introduction

About 35% of Pheochromocytoma (PCC) and Paraganglioma (PGL) patients present with a pathogenic germline or mosaic variant conferring susceptibility to the disease [1]. A total of eighteen genes have been described as disease causing and these loci constitute 25 kilo base pairs spanning through 196 different exons [2–15]. Given the performance of available methods for diagnostic genetic screening, a comprehensive analysis including all PCC and PGL disease causing loci is both extensively recourse demanding and time consuming [16]. Instead, selected fragments are prioritized for diagnostic analysis guided by patient phenotype and/or immunohistochemistry [17]. Such selective screening may have reduced sensitivity and not all patients will be referred for a genetic consultation [18].

Recent publications described next generation sequencing (NGS) workflows for the analysis of genes conferring susceptibility to PCC and PGL [19–21]. Rattenberry et al. suggested near equal quality to Sanger sequencing (SS) and a significant reduction in both cost and time consumption [19]. Similar performance of diagnostic targeted NGS has been reported by an accumulating number of observations in other disease contexts that have used multiple different enrichment assays and sequencing platforms [19, 22–25]. While the robustness of basecalling has been demonstrated across principally different technologies, the performance of multiplexed targeted enrichment and bioinformatic processing have not been thoroughly validated in diagnostic application [26, 27]. Current guidelines for the diagnostic use of next generation sequencing state that the validity of the selected bioinformatic software needs to be ensured by the local investigator [28]. Hence, the local laboratory should select, validate and maintain a robust bioinformatics pipeline, a process that will require trained and experienced personnel. These investments and the running costs of bioinformatic processing will inevitably increase cost of targeted NGS and has been predicted to exceed the total cost of sequencing and enrichment [29].

As current methods impose restrictions in the genetic screening of PCC and PGL patients we initiated a study investigating the use of targeted DNA enrichment, sequenced on a next generation bench top sequencer, utilizing three different bioinformatics tools and compared to results from traditional Sanger sequencing.

Methods

Patients

This was a retrospective study including 21 patients with PCC recruited at two different centres; Uppsala university hospital, Uppsala, Sweden and Kliniken Essen-Mitte, Academic Hospital of the University of Duisburg-Essen, Essen, Germany. Inclusion criteria were; (1) histopathological diagnosis of PCC or PGL; and (2) a confirmed pathogenic genetic mutation

and/or clinical criteria of a PCC or PGL syndrome diagnosed by a specialist in clinical genetics. Of the included patients 17 had a described variant in *SDHB*, *VHL*, *EPAS1* or *RET* and four had clinical criteria of NF1 syndrome [30]. For the Uppsala cohort sequencing data and the presented variants have been partially exploited in previous studies [30–32].

Ethical statement

Samples were obtained from Uppsala Biobank, Endocrine tumour collection (Ethical approval 00-128/ 3.15.2000, Local ethical vetting board in Uppsala (Regionala etikprövningsnämnden i Uppsala)). The study was approved by the regional ethical review board in Uppsala (12-422/ 11.21.2012 and 05-198/ 8.10.2005, Local ethical vetting board in Uppsala (Regionala etikpröv-ningsnämnden i Uppsala)). Written informed consent was obtained from the individual patients. All patients were above 18 years of age at the time of inclusion.

DNA and RNA extraction

Genomic DNA from available tissue samples were extracted using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) as previously described [33]. DNA quality and concentrations were assessed using Nanodrop spectrophotometer (ThermoFischer Scientific, MA, USA) and Qubit Flourometer (Invitrogen, CA, USA). Sample inclusion criteria were a 260/280 spectrums ratio of >1.8 and double strand DNA concentrations above 5ng/μl. RNA was extracted using AllPrep DNA/RNA kit (Qiagen) and was subjected to reverse transcription using RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific, Waltham, MA, USA).

Sanger Sequencing characterization

All included samples had been screened with automated Sanger Sequencing (SS, Beckman Coulter Genomics, Takeley, UK) for a comprehensive panel of 9 disease causing genes to be utilized as negative and positive control (cumulative size 5683 basepairs); *SDHB*, *SDHC*, *SDHD*, *SDHAF2*, *VHL*, *EPAS1* (Exons 9 and 12) *RET* (exons 10–11 and 13–16), *TMEM127* and *MAX* as previously described [30].

Targeted genomic capture

A Truseq Custom Amplicon (Illumina Inc, San Diego, CA, USA) targeted capture and paired end library kit was designed using Illumina Design Studio (Version 2012-09-13, http://designstudio.illumina.com). All coding sequences of eleven established PCC and PGL disease causing genes; *SDHA*, *SDHB*, *SDHC*, *SDHD*, *SDHAF2*, *VHL*, *EPAS1*, *NF1*, *RET* (exons 8, 10–11 and 13–16), *TMEM127* and *MAX* were selected. A novel disease causing gene, *H-RAS* was selected for enrichment but excluded from further analysis because of the lack of clinical relevance for investigating this loci (Table 1). In order to be able to detect variants causing alternative splicing, coordinates were extended with a padding of 10 base pairs at intron-exon boundaries. Coordinates were obtained from the human reference sequence HG19 and the cumulative target size was 24,293 base pairs. The final TruSeq Custom Amplicon design constituted 331 amplicons having a median size of 177 bases. The *in silico* amplicon coverage was >99% with a total gap distance of 57 bases located in a region with homologous sequences; *SDHC* exon 2.

Library preparation and MiSEQ sequencing

Truseq custom amplicon sample kit (Cat. No. FC-130-1001, FC-130-1006, FC-130-1007, Illumina Inc) for targeted capture and library preparation were prepared from 250ng of double

Table 1. Targeted enrichment, included genomic regions.

Gene	Chr	Gene Cordinates	Exons	Cumulative amplicon size, basepairs
SDHA	5	218356–256815	15	6475
SDHB	1	17345217–17380665	8	2625
SDHC	1	161284047–161332984	6	1925
SDHD	11	111957497–111990353	4	1400
SDHAF2	11	61197514–61215001	5	1750
VHL	3	10182692–10193904	3	1400
EPAS1	2	46520806–46613836	16	7700
RET	10	43572475–43625799	20	2800
NF1	17	29421945–29709134	58	27300
TMEM127	2	46520806–46613836	4	2100
MAX	14	65472892–65569413	5	1575
H-RAS	11	532243–535550	6	1050

Selected genomic regions for targeted enrichment utilized in the TruSeq Custom Amplicon assay. Sequences annotated as protein coding were selected in *SDHA, SDHB, SDHC, SDHD, SDHAF2, VHL, EPAS1, RET, NF1, TMEM127, MAX* and *H-RAS*. In *RET* exons 8,10–11,13–16 were selected. The design was extended with 10 base pairs at exon-intron junctions.

doi:10.1371/journal.pone.0133210.t001

stranded DNA accordingly to manufacturers instructions (Part# 15027983). Briefly, upstream and downstream oligonucleotides were hybridized to genomic DNA and unbound oligonucleotides were washed away using ELM3, SW1 and UB1 washing reagents. This was followed by an extension ligation process that connected hybridized upstream and downstream oligonucleotides by using DNA ligase. Extension-ligation products were amplified by PCR and fitted with index adaptor sequences for sample multiplexing using the TruSeq Custom Amplicon Index Kit (Cat. No. FC-130-1003, Illumina Inc). The PCR product was purified from reaction components using AMPure XP beads (Illumina Inc) and selected test samples were run on a 4% agarose gel to confirm successful library preparation. Each library sample underwent quantity normalization by LNA1/LNB1 beads (Illumina Inc) and all 21 samples were pooled together with 74 samples of other origin into a single suspension. Generated paired end (2x150bp) libraries were subjected to two independent sequencing runs on the Illumina MiSEQ platform (Illumina Inc). Sequencing was performed at the university core facility (http://molmed.medsci.uu.se/SNP+SEQ+Technology+Platform/) as instructed in awere automatically demultiplexed by MiSEQ integrated software and results written to .FASTQ files.

Read mapping and variant calling

Generated sequences were processed in-house using three different bioinformatics software workflows (Fig 1): (1) automated pipeline, MiSEQ reporter v2.1.43 (Illumina Inc); (2) semi automated pipeline, CLC Genomics Workbench 5.51 (CLC bio, Aarhus, Denmark); and (3) ICP a in-house custom pipeline constituting of multiple different publicly available software packages as described below. Default settings were used in workflow (1). In workflows (2) and (3) the settings of the variant callers were optimized by using the known pathogenic genetic variants as reference material. The probability and variant allele thresholds were lowered as to achieve the maximum detection of true positive variants in CLC and Freebayes respectively.

In workflow 1 (MSR) a manifest file stating the sequence of hybridizing oligos and the coordinates of attempted amplicons was downloaded from the manufacturer (ww.illumina.com) and uploaded into the MiSEQ reporter 2.1.43 software as instructed (Part# 15038604). Briefly, reads were matched to the sequence coordinates overlapping those of the included truseq
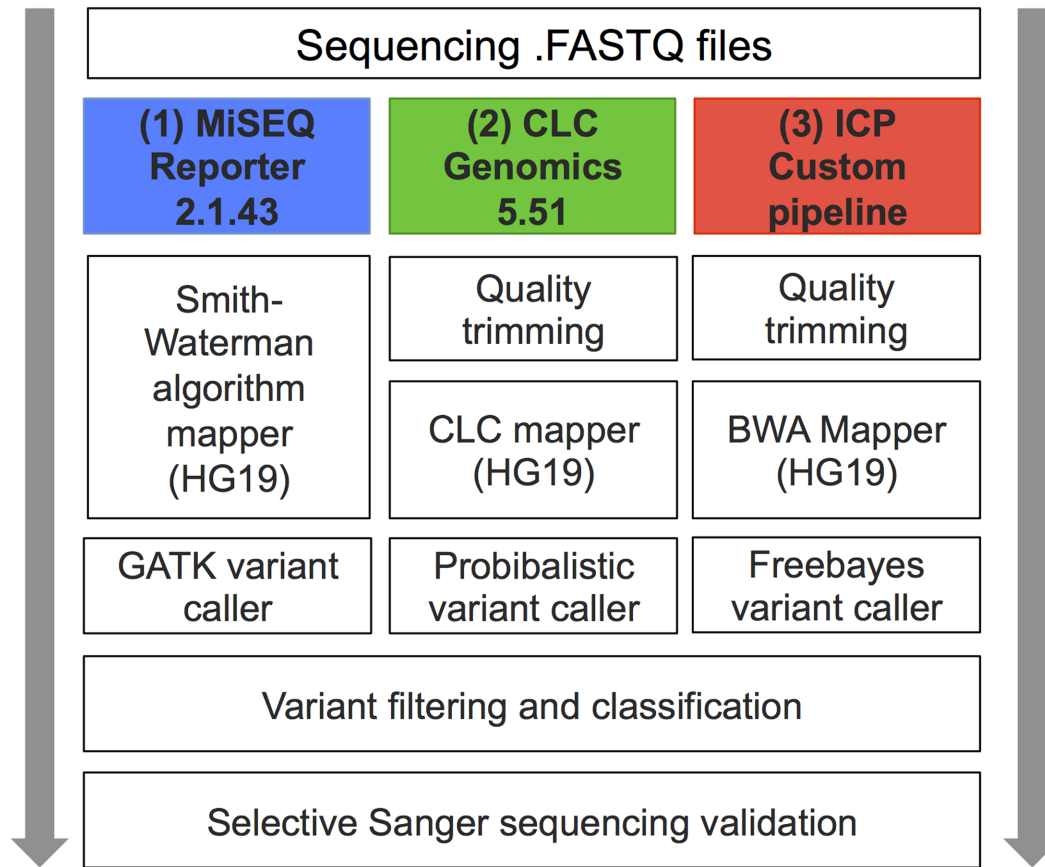
Fig 1. Overview of Bioinformatic workflows. Schematic description of the utilized workflows MiSEQ Reporter 2.1.43 (Illumina inc), CLC Genomics Workbench 5.51 (CLCbio) and the in-house custom pipeline (ICP).

doi:10.1371/journal.pone.0133210.g001

custom amplicon probes. The reads were subsequently aligned to the reference sequence (human reference sequence HG19) using Smith-Waterman algorithm with default settings. Reads that were not matched to probes or having multiple alignments were discarded. Variant calling was performed using a variant caller from the Genome Analysis Toolkit (GATK) with default settings [34]. Generated .BAM and .VCF files were exported for annotation, filtering and further analysis.

In workflow 2 (CLC) the generated .FASTQ files were imported into CLC Genomics Workbench 5.51 and processed with the following steps: 1; Sequence quality trimming based on Phred quality scores and removal of ambiguous base-calls, 2; Read Mapping (human reference sequence HG19) with <3 mismatches / 100bp for alignment, 3; variant calling using a probabilistic based algorithm (Probabilistic Variant Detection Plug-in Manual, clcbio.com) reporting variants having a probability of above 90%. The variant caller was set to exclude variants available in broken read pairs, unspecific read alignments as well as variants exclusively in found forward/reverse reads.

In workflow 3 (ICP) read mapping was performed using the Burrows Wheeler Alignment tool (BWA) with default parameters [35]. The generated SAM files were converted to BAM format, sorted and indexed using Samtools [36]. Variant calling was performed using Freebayes [37]. Variants with a minimum coverage of 30 reads and allele frequency of > 10% were reported. Generated .BAM and .VCF files were exported for annotation, filtering and further analysis.

Generated .BAM files from workflows (1), (2) and (3) were inported into CLC Genomics Workbench 5.51 and analysed for sequence coverage and depth. Targeted bases were defined as the protein coding sequences of the included 11 genes with clinical relevance (excluding *H-RAS*) cumulative size 18324. Generated .VCF files from workflows (1), (2) and (3) were filtered and annotated in CLC Genomics Workbench 5.51. Step 1; removal of synonymous variants without a probable splice site effect. Step 2; the remaining variants were annotated for overlapping information in selected genetic databases; Single Nucleotide Polymorphism database (dbSNP) build 137, Catalogue of Somatic Mutations in Cancer (COSMIC) [38], database of Human Gene Mutation Data (HGMD) [39] and Leiden Open (source) Variant Database (LOVD).The impact of non-synonymous amino acid substitution was assessed *in silico* using Polymorfism Phenotyping v2 (Polyphen2) [40] and Sorting Intolerant from Tolerant (SIFT) [41]. Overlapping variants were analysed with a custom R script. Variants were classified as Pathogenic, Unknown (Variant of Unknown Significance, VUS) or Polymorphism and selected entries were validated with Sanger sequencing. Primer sequences can be obtained by request.

Results

A summary of patient characteristics and discovered genetic variants is presented in Table 2. Samples generated a total of 4961465 (run 01) and 4914971 (run 02) reads. The mean sequencing output per sample was 236380 (range 163818–291262) for run 01 and 230795 (range 163548–302238) for run 02. Results will be presented in parallel for the three workflows; MiSEQ Reporter 2.1.43 (MSR), CLC Genomics Workbench 5.51 (CLC) and the in-house custom pipeline (ICP).

Table 2. Detected mutations and patient characteristics

Patient no.	Age, years	Gender	Syndrome diagnosis	Genotyping
1	15	Male	PGL4	*SDHB* p.Arg90*
2	48	Male	VHL	*VHL* p.Tyr98His
3	26	Male	VHL	*VHL* p.Arg107Ser
4	66	Female	VHL	*VHL* p.Arg161Gln
5	11	Male	VHL	*VHL* p.Val170Gly
6	46	Male	VHL	*VHL* p.*214Gly
7	64	Female	-	*EPAS1* p.Leu529Pro (S)
8	81	Female	-	*EPAS1* p.Ala530Val (S)
9	27	Female	MEN2A	*RET* p.Cys609Ser
10	36	Female	MEN2A	*RET* p.Cys611Tyr
11	29	Female	MEN2A	*RET* p.Cys634Arg
12	29	Male	MEN2A	*RET* p.Cys634Gly
13	30	Female	MEN2A	*RET* p.Cys634Tyr
14	27	Female	MEN2A	*RET* p.Tyr791Phe
15	65	Male	MEN2A	*RET* p.Met804Val
16	34	Male	MEN2B	*RET* p.Met918Leu
17	18	Female	MEN2B	*RET* p.Met918Leu
18	50	Male	NF1	*NF1* p.Arg1241*
19	44	Male	NF1	*NF1* p.Trp2494*
20	57	Female	NF1	*NF1* c.288+1G>T
21	56	Female	NF1	*NF1* p.Lys1844*

Summary of patient characteristics, syndrome criteria and discovered genetic variants. PGL4; Familial Paraganglioma type 4, VHL; Von Hippel Lindau, MEN2; Multiple Endocrine Neoplasia Type 2; and NF1, Neurofibromatosis type 1. (S); Not detected in DNA from peripheral blood.
* indicates stop codon.

doi:10.1371/journal.pone.0133210.t002

Read mapping

Results from read mapping are presented in detail in Table 3 and Figs 2 and 3. Alignment resulted in a mean 87.8% (MSR), 85.5% (CLC) and 89.5% (ICP) of reads mapped to the human reference sequence. The specificities for targeted regions were 52,9% (MSR), 57% (CLC) and 58% (ICP). A total of 97.6% (MSR), 97.8% (CLC) and 98.0% (ICP) of targeted bases had a 10 fold sequencing depth. A 30-fold sequencing depth were achieved in 96.1% (MSR), 97.6% (CLC) and 97.15% (ICP) of the targeted bases. The mean cumulative sizes of targeted regions without mapped reads were 114 (MSR), 91 (CLC) and 92 (ICP) base pairs. The regions without aligned reads were located in *SDHC* exon 2, *SDHAF2* exon 2 and *NF1* exon 57. Alignments of reads to sequences outside the targeted regions were detected in all three pipelines included significant alignment to *SDHA* (*SDHAP1*, *SDHAP2*) and *NF1* (*NF1P1*, *NF1P3*, *NF1P6* and *NF1P8*) psueodogenes.
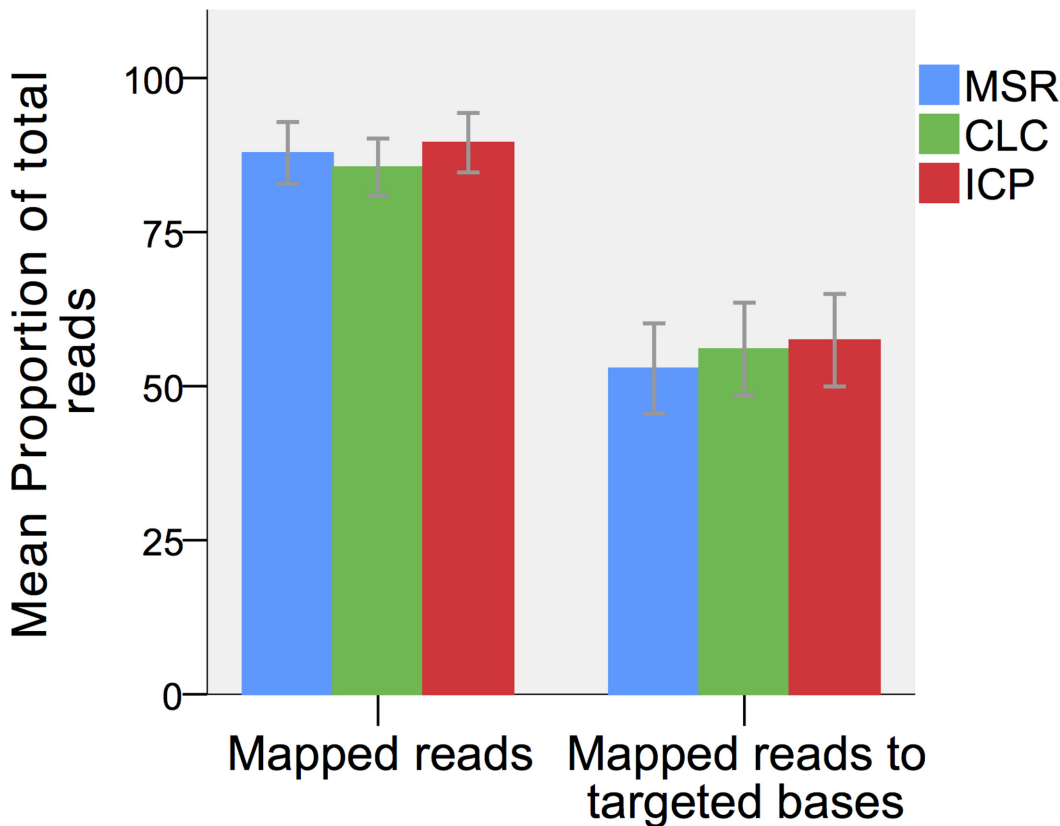
Variant calling

Results from variant calling are presented in detail in Table 4 and Figs 4 and 5. Variant calling revealed a total of 1525 (MSR, Run01; 1418, Run02; 1409), 768 (CLC Run01; 740, Run02; 738) and 1880 (ICP Run01; 1732, Run02; 1747) variants located in the targeted genes. Subsequent filtering of synonymous variants with no probable splice effect resulted in 321 (MSR), 87 (CLC) and 305 (ICP) remaining variants. Out of 47 variants detected by Sanger sequencing, MSR detected all 47 variants in both sequencing runs, CLC detected 39 (run01) and 40 (run02) variants and ICP detected 42 (run01) and 43 (run02) variants. Results from variant calling corresponded to a sensitivity of 100/100% (Run01/02 MSR), 82,9/85,1% (Run 01/02 CLC) and 89,4/91,4% (Run01/02 ICP). CLC did not detect *VHL* p.Tyr98His (run01, patient 3), *EPAS1* p.Leu529Pro (Run 01 and 02, patient 8), *RET* p.Cys611Tyr (Run01, patient 11) and *NF1* p.Arg1241* (run01, patient 19). *SDHA* p.Tyr629Phe was not detected by CLC or freebayse in any

Table 3. Read Mapping

	Mapped reads to HG19 reference, mean % (range)	Mapped reads to targeted regions, mean % (range)	Read coverage, mean (Range)	% X-fold coverage at targeted regions		
				1X, mean (range)	10X, mean (range)	30X, mean (range)
MiSEQ Reporter 2.1.43						
Run01	89.4 (82–92.2)	54 (48.3–64.1)	414 (284–565)	98.1 (97.6–98.4)	97.6 (97.4–98)	95.9 (94.6–96.9)
Run02	86.3 (82.1–88.5)	51.7 (46.6–61.4)	407 (283–571)	98.2 (97.9–98.5)	97.6 (97.3–98.3)	96.2 (94.6–97.3)
CLC Genomics Workbench 5.51						
Run01	86.9 (82.6–92)	56.9 (50.5–67.2)	552 (377–738)	98.4 (98–98.7)	97.9 (97.7–98.4)	97,6 (97,3–97,7)
Run02	84.1 (80–89.1)	55.1 (49.3–64.7)	526 (357–735)	98.5 (98–98.7)	97.8 (97.6–98.7)	97,6 (97,2–98)
In-house custom pipeline						
Run01	89.6 (82–92.2)	58 (51.5–68)	519 (357–699)	98.4 (98–98.7)	98 (97.9–98.3)	97 (96.2–97.6)
Run02	89.4 (82.3–92.4)	56.9 (50.4–66.1)	517 (355–711)	98.5 (98.2–98.7)	98.1 (97.8–98.5)	97.3 (96.3–98)

Results from read mapping. Data presented from three bioinformatics workflows; MiSEQ Reporter 2.1.43 (Smith Waterman algorithm mapper), CLC Genomics Workbench 5.51 (default CLC mapper) and in-house custom pipeline (Burrows Wheeler Alignment tool).

doi:10.1371/journal.pone.0133210.t003

Fig 2. Read mapping. Results from the three different workflows; MSR; MiSEQ Reporter 2.1.43 (Illumina Inc), CLC; CLC Genomics Workbench 5.51 (CLCbio); and ICP, in-house custom pipeline. The proportion of reads mapped to the reference sequence and targeted bases respectively.

doi:10.1371/journal.pone.0133210.g002

of the sequencing runs. The specificity was >99.99% for MSR and ICP while CLC had a perfect 100% specificity (Table 5). The number of false positive variants could be reduced by removal of variants not available in both sequencing runs in the MSR and ICP workflows. In total 17% of variants were reported among all workflows and about 60% were specific to a single workflow.

Variant classification

All pathogenic variants occurred in a non-concomitant fashion and all but one patient had a pathogenic variant in one of the investigated genes. One patient had a pathogenic germline mutation in *SDHB* (p.Arg90*). Five different pathogenic germline variants were confirmed in patients with clinical criteria of von Hippel Lindau syndrome (p.Tyr98His, p.Arg107Ser, p. Arg161Gln, p.Val170Gly and p.*214G). There were two pathogenic *EPAS1* variants (p.Leu529-Pro and p.Ala530Val) that were detected in DNA from tumoral tissue, these variants were absent in DNA from peripheral blood. Eight patients with clinical criteria of MEN2 had pathogenic germline variants in *RET* (p.Cys609Ser, p.Cys611Tyr, p.Cys634Arg, p.Cys634Gly, p. Cys634Tyr, p.Met804Val, p.Met918Leu and p.Met918Leu). There were three nonsense variants detected in germline DNA from three different patients with clinical criteria of NF1 syndrome; p.Arg1241*, p.Lys1844* and p.Trp2494* (Fig 6). One additional patient with NF1 syndrome had a germline variant resulting in alternative splicing; c.288+1G>T. Sequencing of cDNA
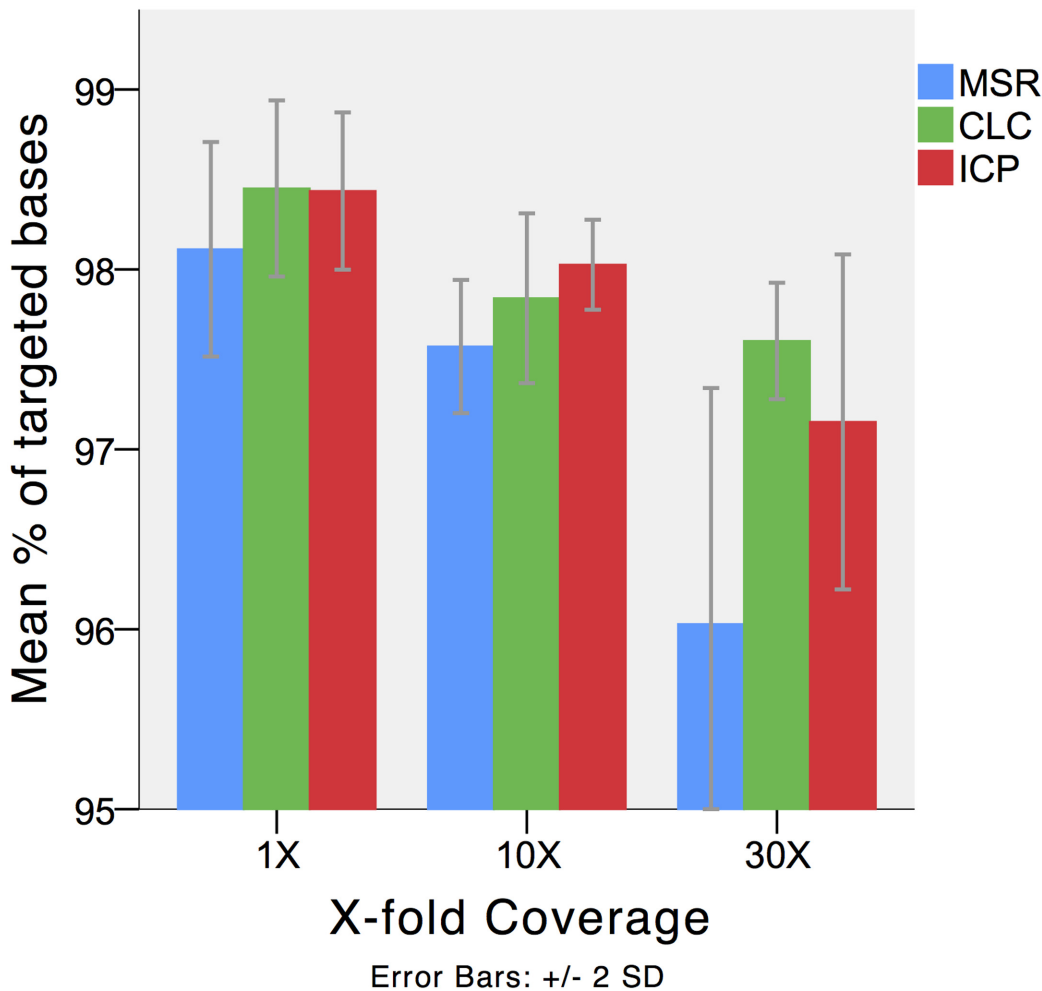
Fig 3. Coverage. Results from the three different workflows; MSR; MiSEQ Reporter 2.1.43 (Illumina Inc), CLC; CLC Genomics Workbench 5.51 (CLCbio); and ICP, in-house custom pipeline. The proportion of targeted reads that achieved X-fold coverage. Error bars represent +/- 2 standard deviations.

doi:10.1371/journal.pone.0133210.g003

derived from peripheral blood revealed skipping of exon 4. These four variants in *NF1* were all classified as pathogenic. One patient had only variants of unknown significance including *RET* p.Tyr791Phe.

Performance

Following implementation of the workflow and optimization of the bioinformatics workflow, the total throughput time for NGS was 7 days divided on 1 day for sample preparation and quality assessment, 2 days for sample enrichment and multiplexing, 1 day for MiSEQ sequencing and 1 day for bioinformatics processing and interpretation. Validation with Sanger sequencing may be estimated to an additional 7–14 days. In our laboratory we were able to reduce the cost of sequencing per exon from 6.5 USD to 0.56 USD for this experiment (S1 Table).

Discussion

The present study validated an amplicon based next generation sequencing method for diagnostic re-sequencing of 11 genes (including *EPAS1* and *NF1*) associated with PCC and PGL

Table 4. Variant calling

	Total variants	Variants per sample, mean (range)	Substitutions per sample, mean (range)	INDELs per sample, mean (range)
MiSEQ Reporter 2.1.43				
Run01	1418	75 (30–92)	63 (23–80)	12 (7–15)
Run02	1409	75 (60–96)	62 (48–82)	12 (6–15)
CLC Genomics Workbench 5.51				
Run01	740	34 (22–52)	28 (17–44)	6 (3–9)
Run02	738	32 (20–51)	27 (17–42)	5 (3–9)
In-house custom pipeline				
Run01	1732	85 (40–102)	73 (33–90)	12 (7–15)
Run02	1747	85 (70–106)	72 (58–92)	12 (6–15)

Results from variant calling in the three workflows MiSEQ Reporter 2.1.43 (GATK variant caller), CLC Genomics Workbench 5.51 (Probalistic variant caller) and the in-house custom pipeline (Freebayse variant caller). Results are presented as total number of variants in targeted regions and variants per sample.

doi:10.1371/journal.pone.0133210.t004

tumours. To determine the robustness of bioinformatic processing, three principally different bioinformatics pipelines were compared. Only the fully automated and integrated software package reported all variants detected by Sanger Sequencing. We confirmed that targeted NGS have superior performance and comparable quality to Sanger Sequencing. Our results further suggest that the bioinformatics analysis needs to be carefully reviewed before clinical application and that the analysis can be performed using automated software.

The mapping process generated comparable results across the different workflows with about 85–90% and 55% of reads being mapped to the human reference and targeted sequences respectively. Several factors may contribute to this relatively low on target proportion. Multiple amplicons were located at intro/exon boundaries, and as a consequence a proportion of sequencing reads were located outside the protein coding sequences (targeted regions).
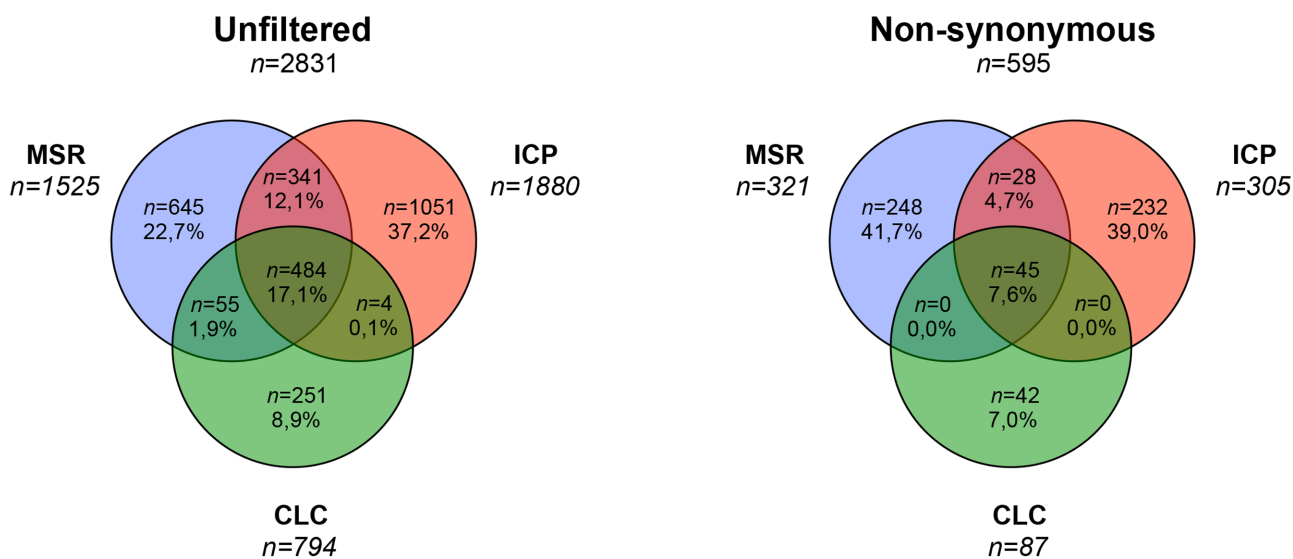

Fig 4. Venn diagram of overlapping variants between the three workflows, total (all variants available at bases annotated for the 11 included genes) and non synonymous remaining variants after filtering synonymous variants with no calculated splice site disruption. MSR; MiSEQ Reporter, CLC; CLC Genomics Workbench, and ICP; In-house custom pipeline.

doi:10.1371/journal.pone.0133210.g004

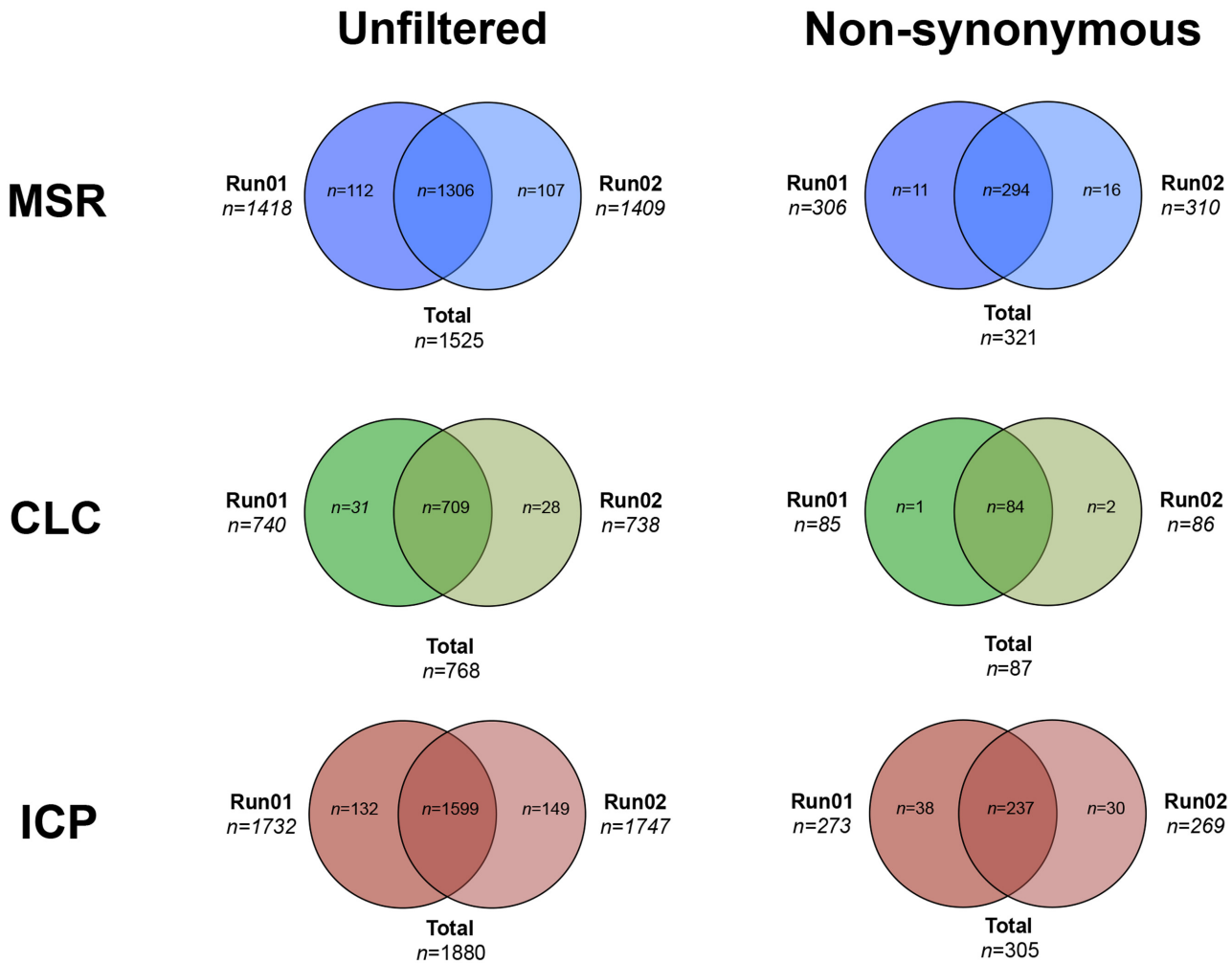## Unfiltered

Non-synonymous

Fig 5. Venn diagram of overlapping variants between the two sequencing runs, total (all variants available at bases annotated for the 11 included genes) and non synonymous remaining variants after filtering synonymous variants with no calculated splice site disruption. MSR; MiSEQ Reporter, CLC; CLC Genomics Workbench, and ICP; In-house custom pipeline.

Unspecific read alignment was also detected with reads mapped to *SDHA* and *NF1* pseudogenes as well as to genomic regions without annotation. The proportion of targeted sequences that achieved an appropriate read coverage was similar to previous studies that have identified regions with high GC content being difficult to amplify during enrichment and library preparation [21, 42]. Indeed several of the genes conferring susceptibility to PCC and PGL have high GC content and/or multiple pseudogenes that may complicate the design and interpretation of genetic testing and 100% coverage may be hard to achieve [19]. A potential impact of the unspecific amplification cannot be ruled out as a defined region in *NF1* exon 21 had a relatively high number of false positive variants. But as these findings occurred in a stochastic manner and the variants could be removed by subtracting variants not available in both sequencing runs. To reduce unspecific amplification, a high degree of flexibility regarding amplicon size and oligo location is warranted when designing the multiplex assay for genomic enrichment.

With regards to variant calling there was pronounced heterogeneity observed between the three workflows. Only a minority of the detected variants were shared between the three bioinformatics pipelines. Examining these variants that were not overlapping between the workflows

Table 5. Sensitivity and specificity compared to Sanger Sequencing.

	True pos, n	True neg, n	False pos, n	False neg, n	Sensitivity, %	Specificity, %
MiSEQ Reporter 2.1.43						
Run01	47	119197	99	0	100	>99.99
Run02	47	119195	101	0	100	>99.99
Filtered, merged Runs 01+02	47	119190	94	0	100	>99.99
CLC Genomics Workbench 5.51						
Run01	39	119296	0	8	82.9	100
Run02	40	119296	0	7	85.1	100
Filtered, merged Runs 01+02	36	119296	0	11	76.7	100
In-house Custom pipeline						
Run01	42	119284	7	5	89.4	>99.99
Run02	43	119288	4	4	91.4	>99.99
Filtered, merged Runs 01+02	42	119285	4	7	87.2	>99.99

Sensitivity of targeted next generation sequencing compared to current golden standard (automated Sanger sequencing) covering 5683 basepairs. Results are presented separately for both sequencing runs as well. Filtered and merged results incudes only variants available in both sequencing runs. Pos; positive, neg; negative.

doi:10.1371/journal.pone.0133210.t005

were most often positioned outside protein coding regions in amplicon start/ends and often occurred in a high frequency of the included cases. Only one workflow managed to detect all variants within the reference panel; MSR. The MSR analysis was performed with default settings that achieved optimal processing despite the limited flexibility of the software. The ICP workflow failed to detect one pathogenic variant, located to *RET* codon 804, in one of the sequencing runs. This false negative was probably due to low coverage, a phenomenon that was detected in this region in all three workflows. CLC did not detect several variants despite extensive optimization with focus on sensitivity for pathogenic variants. Even so this workflow generated the lowest number of variants and had the highest specificity and it cannot be ruled out that the selected settings for variant calling was to stringent resulting in a lower sensitivity. The variability of generated variants among the different workflows is comparable to previous studies showing similar differences between NGS bioinformatic pipelines [43, 44].

The rationale to include workflows having a graphical interface and a high degree of automatization in diagnostic bioinformatics analysis would be potential cost reductions. This may be achieved through outsourcing of certain bioinformatics tasks to staff with intermediate
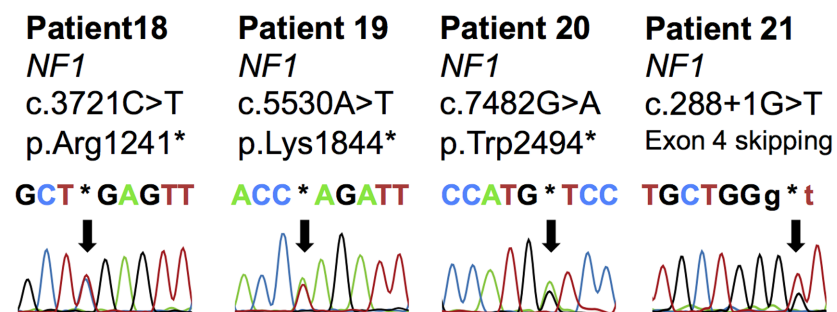

Fig 6. Detected NF1 mutations. Chromatograms from automated Sanger sequencing as displayed in CLC Genomics Workbench 5.51 (CLCbio). *NF1* variants available in germline DNA that were classified as pathogenic.

doi:10.1371/journal.pone.0133210.g006

computational skills. Due to the full integration of the MiSEQ reporter software, the total hands-on time was reduced to a few minutes and there was no time needed for optimization of the workflow. A graphical interface and high degree of automatization is shared by CLC that allows for a higher degree of flexibility in both read mapping and variant detection. However, the optimization process was long, as the workflow could not be tuned to report all variants detected by Sanger sequencing. The command-based workflow (ICP) had an intermediate profile both in regards of performance and total hands on time. Our results suggest that software with a graphical interface and a high degree of automatization may allow outsourcing of certain tasks to less experienced staff and could therefore be cost effective (with equal quality).

The momentum of NGS in a clinical setting was recently strengthened by demonstrating equal quality of generated results compared to SS [45]. A subsequent proof of principle study for the analysis of nine genes associated with PCC and PGL tumours suggested that targeted next generation sequencing would be beneficial with a 70% cost reduction and 66% increase in diagnostic yield compared to sanger sequencing [19]. Results from this study confirmed these specifications and were further able to screen the *NF1* gene and *EPAS1* for somatic mutations. Germline mutations in *NF1* and mosaic mutations in *EPAS1* have recently been found in apparently sporadic PCC patients, and would increase the diagnostic yield if in included into the analysis [21, 46, 47].

We conclude that targeted next generation sequencing has equal quality compared to Sanger sequencing. Enrichment specificity and the bioinformatic sensitivity need to be ensured in each clinical diagnostic application. As superior accuracy was noted for a fully automated bioinformatic workflow compared to two other bioinformatics tools, we suggest that handling of NGS data could be performed without expert bioinformatics skills utilizing commercially available software.

Supporting Information

S1 Table. Cost for bioinformatic workflows. * Given that sequencing provider provides the analysis.
(DOCX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JC PB. Performed the experiments: JC VL. Analyzed the data: JC VL PB. Contributed reagents/materials/analysis tools: SW MKW PH PB. Wrote the paper: JC PB.

References

1. Martucci VL, Pacak K. Pheochromocytoma and paraganglioma: diagnosis, genetics, management, and treatment. Current problems in cancer. 2014; 38(1):7–41. Epub 2014/03/19. doi: 10.1016/j.currproblcancer.2014.01.001 PMID: 24636754; PubMed Central PMCID: PMCPmc3992879.

2. Astuti D, Latif F, Dallol A, Dahia PL, Douglas F, George E, et al. Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. Am J Hum Genet. 2001; 69(1):49–54. Epub 2001 Jun 12. PMID: 11404820

3.   Niemann S, Muller U. Mutations in SDHC cause autosomal dominant paraganglioma, type 3. Nat Genet. 2000; 26(3):268–70. PMID: 11062460

4.   Burnichon N, Briere JJ, Libe R, Vescovo L, Riviere J, Tissier F, et al. SDHA is a tumor suppressor gene causing paraganglioma. Hum Mol Genet. 2010; 19(15):3011–20. Epub 2010/05/21. ddq206 [pii] doi: 10.1093/hmg/ddq206 [doi]. PMID: 20484225; PubMed Central PMCID: PMC2901140.

5.   Baysal BE, Ferrell RE, Willett-Brozick JE, Lawrence EC, Myssiorek D, Bosch A, et al. Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. Science. 2000; 287(5454):848–51. PMID: 10657297

6.   Bayley JP, Kunst HP, Cascon A, Sampietro ML, Gaal J, Korpershoek E, et al. SDHAF2 mutations in familial and sporadic paraganglioma and phaeochromocytoma. Lancet Oncol. 2010; 11(4):366–72. Epub 2010 Jan 11. doi: 10.1016/S1470-2045(10)70007-3 PMID: 20071235

7.   Latif F, Tory K, Gnarra J, Yao M, Duh FM, Orcutt ML, et al. Identification of the von Hippel-Lindau disease tumor suppressor gene. Science. 1993; 260(5112):1317–20. PMID: 8493574

8.   Zhuang Z, Yang C, Lorenzo F, Merino M, Fojo T, Kebebew E, et al. Somatic HIF2A gain-of-function mutations in paraganglioma with polycythemia. N Engl J Med. 2012; 367(10):922–30. doi: 10.1056/NEJMoa1205119 PMID: 22931260

9.   Wallace MR, Marchuk DA, Andersen LB, Letcher R, Odeh HM, Saulino AM, et al. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. Science. 1990; 249 (4965):181–6. PMID: 2134734

10. Donis-Keller H, Dou S, Chi D, Carlson KM, Toshima K, Lairmore TC, et al. Mutations in the RET proto-oncogene are associated with MEN 2A and FMTC. Hum Mol Genet. 1993; 2(7):851–6. PMID: 8103403

11. Qin Y, Yao L, King EE, Buddavarapu K, Lenci RE, Chocron ES, et al. Germline mutations in TMEM127 confer susceptibility to pheochromocytoma. Nat Genet. 2010; 42(3):229–33. Epub 2010 Feb 14. doi: 10.1038/ng.533 PMID: 20154675

12. Comino-Mendez I, Gracia-Aznarez FJ, Schiavi F, Landa I, Leandro-Garcia LJ, Leton R, et al. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. Nat Genet. 2011; 43(7):663–7. Epub 2011/06/21. ng.861 [pii] doi: 10.1038/ng.861 [doi]. PMID: 21685915.

13. Letouze E, Martinelli C, Loriot C, Burnichon N, Abermil N, Ottolenghi C, et al. SDH Mutations Establish a Hypermethylator Phenotype in Paraganglioma. Cancer Cell. 2013; 21(13):00183–9.

14. Ladroue C, Carcenac R, Leporrier M, Gad S, Le Hello C, Galateau-Salle F, et al. PHD2 mutation and congenital erythrocytosis with paraganglioma. N Engl J Med. 2008; 359(25):2685–92. doi: 10.1056/NEJMoa0806277 PMID: 19092153

15. Crona J, Delgado Verdugo A, Maharjan R, Stalberg P, Granberg D, Hellman P, et al. Somatic mutations in H-RAS in sporadic pheochromocytoma and paraganglioma identified by exome sequencing. J Clin Endocrinol Metab. 2013; 98(7):E1266–71. Epub 2013/05/04. doi: 10.1210/jc.2012-4257 PMID: 23640968.

16. Erlic Z, Rybicki L, Peczkowska M, Golcher H, Kann PH, Brauckhoff M, et al. Clinical predictors and algorithm for the genetic diagnosis of pheochromocytoma patients. Clin Cancer Res. 2009; 15 (20):6378–85. Epub 2009/10/15. doi: 10.1158/1078-0432.ccr-09-1237 PMID: 19825962.

17. Jafri M, Maher ER. The genetics of phaeochromocytoma: using clinical features to guide genetic testing. Eur J Endocrinol. 2012; 166(2):151–8. doi: 10.1530/EJE-11-0497 Epub 2011 Sep 6. PMID: 21896620

18. Jafri M, Whitworth J, Rattenberry E, Vialard L, Kilby G, Kumar AV, et al. Evaluation of SDHB, SDHD and VHL gene susceptibility testing in the assessment of individuals with non-syndromic phaeochromocytoma, paraganglioma and head and neck paraganglioma. Clin Endocrinol (Oxf). 2013; 78(6):898–906. doi: 10.1111/cen.12074 Epub 2013 Apr 6.

19. Rattenberry E, Vialard L, Yeung A, Bair H, McKay K, Jafri M, et al. A comprehensive next generation sequencing based genetic testing strategy to improve diagnosis of inherited pheochromocytoma and paraganglioma. J Clin Endocrinol Metab. 2013; 10:10.

20. Casey R, Garrahy A, Tuthill A, O'Halloran D, Joyce C, Casey MB, et al. Universal Genetic Screening Uncovers a Novel Presentation of an SDHAF2 Mutation. J Clin Endocrinol Metab. 2014:jc20134536. Epub 2014/04/10. doi: 10.1210/jc.2013-4536 PMID: 24712571.

21. Welander J, Andreasson A, Juhlin CC, Wiseman RW, Backdahl M, Hoog A, et al. Rare germline mutations identified by targeted next-generation sequencing of susceptibility genes in pheochromocytoma and paraganglioma. J Clin Endocrinol Metab. 2014:jc20134375. Epub 2014/04/04. doi: 10.1210/jc.2013-4375 PMID: 24694336.

22. Chan M, Ji SM, Yeo ZX, Gan L, Yap E, Yap YS, et al. Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. J Mol Diagn. 2012; 14(6):602–12. Epub Aug 22. PMID: 22921312

23. Pritchard CC, Smith C, Salipante SJ, Lee MK, Thornton AM, Nord AS, et al. ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. J Mol Diagn. 2012; 14(4):357–66. doi: 10.1016/j.jmoldx.2012.03.002 Epub May 30. PMID: 22658618

24. Feliubadalo L, Lopez-Doriga A, Castellsague E, Del Valle J, Menendez M, Tornero E, et al. Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. Eur J Hum Genet. 2012; 19(10):270.

25. Halbritter J, Diaz K, Chaki M, Porath JD, Tarrier B, Fu C, et al. High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. J Med Genet. 2012; 49(12):756–67. doi: 10.1136/jmedgenet-2012-100973 PMID: 23188109

26. Grossmann V, Roller A, Klein HU, Weissmann S, Kern W, Haferlach C, et al. Robustness of Amplicon Deep Sequencing Underlines Its Utility in Clinical Applications. J Mol Diagn. 2013; 13(13):00057–3.

27. Jacob HJ, Abrams K, Bick DP, Brodie K, Dimmock DP, Farrell M, et al. Genomics in clinical practice: lessons from the front lines. Sci Transl Med. 2013; 5(194):194cm5. doi: 10.1126/scitranslmed.3006468 PMID: 23863829

28. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012; 30(11):1033–6. PMID: 23138292

29. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011; 12(8):125. doi: 10.1186/gb-2011-12-8-125 PMID: 21867570

30. Crona J, Nordling M, Maharjan R, Granberg D, Stalberg P, Hellman P, et al. Integrative genetic characterization and phenotype correlations in pheochromocytoma and paraganglioma tumours. PloS one. 2014; 9(1):e86756. Epub 2014/01/28. doi: 10.1371/journal.pone.0086756 PMID: 24466223; PubMed Central PMCID: PMCPmc3899286.

31. Crona J, Backman S, Maharjan R, Mayrhofer M, Stålberg P, Isaksson A, et al. Spatio-temporal heterogeneity characterizes the genetic landscape of pheochromocytoma and defines early events in tumourigenesis. Clinical Cancer Research. 2015.

32. Crona J, Maharjan R, Delgado Verdugo A, Stålberg P, Granberg D, Hellman P, et al. MAX mutations status in Swedish patients with pheochromocytoma and paraganglioma tumours. Fam Cancer. 2013; 7:7.

33. Åkerstrom T, Crona J, Delgado Verdugo A, Starker LF, Cupisti K, Willenberg HS, et al. Comprehensive re-sequencing of adrenal aldosterone producing lesions reveal three somatic mutations near the KCNJ5 potassium channel selectivity filter. PLoS One. 2012; 7(7):e41926. doi: 10.1371/journal.pone.0041926 Epub 2012 Jul 27. PMID: 22848660

34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. Epub 2010 Jul 19. PMID: 20644199

35. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009 May 18. PMID: 19451168

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. doi: 10.1093/bioinformatics/btp352 Epub 2009 Jun 8. PMID: 19505943

37. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012.

38. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39(Database issue):D945–50. Epub 2010/10/19. doi: 10.1093/nar/gkq929 PMID: 20952405; PubMed Central PMCID: PMC3013785.

39. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003; 21(6):577–81. PMID: 12754702

40. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–9. doi: 10.1038/nmeth0410-248 PMID: 20354512

41. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4(7):1073–81. Epub 2009 Jun 25. doi: 10.1038/nprot.2009.86 PMID: 19561590

42. Li X, Buckton AJ, Wilkinson SL, John S, Walsh R, Novotny T, et al. Towards clinical molecular diagnosis of inherited cardiac conditions: a comparison of bench-top genome DNA sequencers. PLoS One. 2013; 8(7):e67744. doi: 10.1371/journal.pone.0067744 Print 2013. PMID: 23861798

43. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013; 5(3):28. doi: 10.1186/gm432 PMID: 23537139

44. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. PLoS One. 2013; 8(9):e75619. doi: 10.1371/journal.pone.0075619 PMID: 24086590

45. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, et al. Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. Hum Mutat. 2013; 8(10):22332.

46. Burnichon N, Buffet A, Parfait B, Letouze E, Laurendeau I, Loriot C, et al. Somatic NF1 Inactivation is a Frequent Event in Sporadic Pheochromocytoma. Hum Mol Genet. 2012; 6:6.

47. Korf BR. Diagnosis and management of neurofibromatosis type 1. Curr Neurol Neurosci Rep. 2001; 1 (2):162–7. PMID: 11898512