ORIGINAL ARTICLE

Impact of DNA source on genetic variant detection from human whole-genome sequencing data

Brett Trost,[1] Susan Walker,[1] Syed A Haider,[1] Wilson W L Sung,[1] Sergio Pereira,[1] Charly L Phillips,[1] Edward J Higginbotham,[1,2] Lisa J Strug,[1,3] Charlotte Nguyen,[1,2] Akshaya Raajkumar,[1] Michael J Szego,[4,5] Christian R Marshall,[6,7] Stephen W Scherer[1,2]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2019-106281).

[1]The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Ontario, Canada
[2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[3]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada
[4]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[5]Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada
[6]Department of Paediatric Laboratory Medicine, Genome Diagnostics, Hospital for Sick Children, Toronto, Ontario, Canada
[7]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Stephen W Scherer, The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Ontario, Canada; stephen.scherer@sickkids.ca

Received 12 May 2019
Revised 4 July 2019
Accepted 20 July 2019
Published Online First 12 September 2019

To cite: Trost B, Walker S, Haider SA, *et al. J Med Genet* 2019;**56**:809–817.

ABSTRACT

Background Whole blood is currently the most common DNA source for whole-genome sequencing (WGS), but for studies requiring non-invasive collection, self-collection, greater sample stability or additional tissue references, saliva or buccal samples may be preferred. However, the relative quality of sequencing data and accuracy of genetic variant detection from blood-derived, saliva-derived and buccal-derived DNA need to be thoroughly investigated.

Methods Matched blood, saliva and buccal samples from four unrelated individuals were used to compare sequencing metrics and variant-detection accuracy among these DNA sources.

Results We observed significant differences among DNA sources for sequencing quality metrics such as percentage of reads aligned and mean read depth ($p<0.05$). Differences were negligible in the accuracy of detecting short insertions and deletions; however, the false positive rate for single nucleotide variation detection was slightly higher in some saliva and buccal samples. The sensitivity of copy number variant (CNV) detection was up to 25% higher in blood samples, depending on CNV size and type, and appeared to be worse in saliva and buccal samples with high bacterial concentration. We also show that methylation-based enrichment for eukaryotic DNA in saliva and buccal samples increased alignment rates but also reduced read-depth uniformity, hampering CNV detection.

Conclusion For WGS, we recommend using DNA extracted from blood rather than saliva or buccal swabs; if saliva or buccal samples are used, we recommend against using methylation-based eukaryotic DNA enrichment. All data used in this study are available for further open-science investigation.

INTRODUCTION

Whole blood is the most common source of DNA for genetic analyses in both research and clinical settings. This is presumably for historical reasons—early studies of genetic disease used blood-derived DNA,[1] and there exist established procedures and infrastructure for biochemical and metabolite testing in blood. However, blood collection can be problematic, especially for populations without access to phlebotomy centres and for individuals unwilling or unable to give blood.[2] Alternative sources of DNA include saliva and buccal (cheek) cells, which are becoming increasingly popular due to ease of collection (including being non-invasive and amenable to self-collection) and better stability for shipping and storage.[2 3]

Whole-genome sequencing (WGS) is gradually replacing whole-exome sequencing and chromosomal microarray analysis (CMA) for genetic variant detection, since WGS can detect all sizes and types of variants with base-pair resolution in one experiment. However, in order for WGS to achieve the broadest possible impact across precision medicine[4 5] and general biology,[6 7] a better understanding of the impact of DNA source is required. Despite their advantages, saliva and buccal samples will not become equally accepted DNA sources for WGS until all classes of genetic variation can be detected from them as accurately as from blood samples.

Previous studies have compared genetic variant detection from blood-derived DNA to that of DNA isolated from saliva or buccal samples. Most reported no difference in accuracy,[8–19] although some favoured blood-derived DNA[20–22] (online supplementary table 1; all supplementary tables and figures are in online supplementary file 1). However, all but one of these studies used CMA, so their applicability to WGS is unclear. Further, although all prior studies examined single nucleotide polymorphisms (SNPs; single-base substitutions of moderate-to-high population frequency), few assessed copy number variants (CNVs), none examined short insertions/deletions (indels) and just one (the sole WGS study[11]) assessed single nucleotide variants (SNVs; single-base substitutions of any frequency) (online supplementary table 1).

Here, we performed a comprehensive assessment of the impact of DNA source using industry-standard short-read WGS data. Our systematic study design investigated how DNA source and bacterial DNA contamination affect the quality of sequencing data and the accuracy of SNV, indel, and CNV detection. We also investigated a methylation-selection method for reducing bacterial DNA contamination in saliva and buccal samples prior to sequencing.[23] All samples and data were from Personal Genome Project Canada (PGPC)[24] participants, who consented for open sharing.

METHODS

From each of four individuals who had previously provided blood samples for the PGPC study,[24] we collected three saliva samples and three buccal samples (all on different days). Online supplementary table 2 indicates the age of each participant at sample collection. We quantified bacterial DNA for each sample and selected one saliva and one buccal sample per individual for further analysis. DNA library preparation (PCR-free) and sequencing (Illumina HiSeq X) were performed for each blood sample, as well as for each selected saliva and buccal sample either with or without prior methylation-based enrichment for eukaryotic DNA. (Generally, eukaryotic DNA is methylated but microbial DNA is not, allowing separation based on methylation status.[23]) SNVs and indels were detected using the Genome Analysis Toolkit,[25] and CNVs were detected using ERDS[26] and CNVnator[27] as previously described.[28] We then identified differences in sequencing metrics and variant-detection accuracy among the sample types. As a baseline for variant-detection concordance, we used a previously generated sequencing data set from HuRef blood-derived DNA,[28] as well as a second replicate from the same DNA extraction prepared and sequenced specifically for this study. Although this study is largely descriptive, when appropriate we used statistical tests tailored to small sample sizes. The online supplementary file 2 contains full details on DNA extraction, bacterial DNA quantification, eukaryotic DNA enrichment, DNA library preparation and sequencing, variant detection and statistical analysis.

RESULTS

Bacterial DNA quantification

From each of four study participants, denoted PGPC-0002, PGPC-0005, PGPC-0006 and PGPC-0050, we collected one blood sample, three saliva samples and three buccal samples and quantified their relative concentrations of human and bacterial DNA. As expected, the blood samples contained little bacterial DNA (online supplementary figure 1). Generally, there was substantially more bacterial DNA in saliva than in buccal samples, and its concentration varied more in saliva samples both among and within individuals. For further analysis, we selected one saliva and one buccal sample per individual, representing a range of bacterial DNA concentrations (online supplementary figure 1). Five WGS data sets were generated per individual, derived from blood, saliva without eukaryotic DNA enrichment, saliva with enrichment, buccal without enrichment and buccal with enrichment (figure 1). The WGS data sets were then evaluated for general WGS and alignment characteristics and variant-detection concordance and accuracy.

General WGS and alignment characteristics

Statistically significant differences among blood, non-enriched saliva and non-enriched buccal samples were observed for several sequencing metrics (Friedman repeated-measures test, followed by Conover-Iman tests to assess pairwise differences). For example, the percentage of reads successfully aligned to the human reference genome was significantly higher in blood samples (99.8%±0.1%) than in non-enriched saliva samples (85.3%±10.7%; p=0.000 for mean different from blood) and non-enriched buccal samples (98.4%±0.7%; p=0.005) (online supplementary tables 3-4). Blood samples also had significantly lower percentages of alignments <50 bp (typical of bacterial DNA), higher mean sequencing depths and lower mean mitochondrial sequencing depths. The percentage of alignments <50 bp was significantly lower in enriched saliva (0.6%±0.5%) and buccal (0.1%±0.0%) samples than in non-enriched saliva (4.8%±4.3%) and buccal (0.4%±0.2%) samples (Wilcoxon signed-rank test p=0.062 for both saliva and buccal), suggesting that enrichment successfully removed bacterial DNA (online supplementary tables 3-4). Compared with their non-enriched counterparts, the enriched saliva and buccal samples also had significantly higher percentages of aligned reads, higher percentages of genomic positions sequenced to >40× depth and lower mean mitochondrial sequencing depths. Enriched saliva samples also had significantly higher mean genome-wide sequencing depths and higher percentages of genomic positions sequenced to >30× depth than non-enriched saliva samples. Enriched samples exhibited lower read-depth uniformity, particularly for buccal (online supplementary table 3 and online supplementary figure 2).

To determine whether differing bacterial DNA concentrations were driving these observations, we plotted bacterial DNA concentration against each sequencing metric. When non-enriched, the two samples with the highest bacterial DNA concentrations (both saliva; online supplementary figure 1) had the highest percentages of aligned sequences <50 bp and the lowest values for percentage of reads aligned, mean mapping quality, median insert size, mean genome-wide read depth and percentage of genomic positions sequenced to >40× depth (figure 2). When the same samples were enriched, the values of these metrics approached those of the samples with lower bacterial DNA concentrations. Enrichment had a material impact on these sequencing metrics only for samples with high bacterial DNA concentrations.

To determine their sources, we used BLAST to search 10 000 unmapped reads from each sample against the National Center for Biotechnology Information (NCBI) nucleotide database (online supplementary table 5). As expected, the percentage of unmapped reads matching bacteria was highest in the non-enriched saliva and buccal samples, lower in the corresponding enriched samples, and nearly zero in the blood samples. Most unmapped reads from blood matched eukaryotes, suggesting that sequencing errors may explain why they were unmapped. The percentage of unmapped reads in a given sample that matched bacteria was positively correlated with their mean base-quality score (online supplementary table 5), suggesting that unmapped reads not matching bacteria were more likely to arise from sequencing errors.

Impact of DNA source and eukaryotic DNA enrichment on SNV and indel detection

To eliminate mean sequencing depth as a confounding variable, reads were subsampled prior to variant detection to give each sample approximately the same mean depth as the lowest-depth sample (25×). To begin comparing SNV and indel detection among the five sample types, we computed variant counts and allele fraction distributions for each sample. After filtering, counts of known variants (those in the Genome Aggregation Database (gnomAD)[29]) ranged between 3 530 091 and 3 674 442 for SNVs and between 218 964 and 226 099 for indels; counts for novel variants (those absent from gnomAD) ranged between 20 072 and 58 060 for SNVs and between 2950 and 4223 for indels (online supplementary table 6). No statistically significant differences were observed among the five sample types in terms of the number of variants detected in each category (known SNVs, novel SNVs, known indels and novel indels) (Friedman repeated-measures test for blood, non-enriched saliva and non-enriched buccal samples or Wilcoxon signed-rank test for enriched vs non-enriched saliva or buccal samples). Allele fraction distributions did not differ with sample type (online supplementary figure 3 and online supplementary table 7).

Next, we compared blood-derived DNA with DNA from non-enriched saliva and buccal samples in terms of SNV and indel

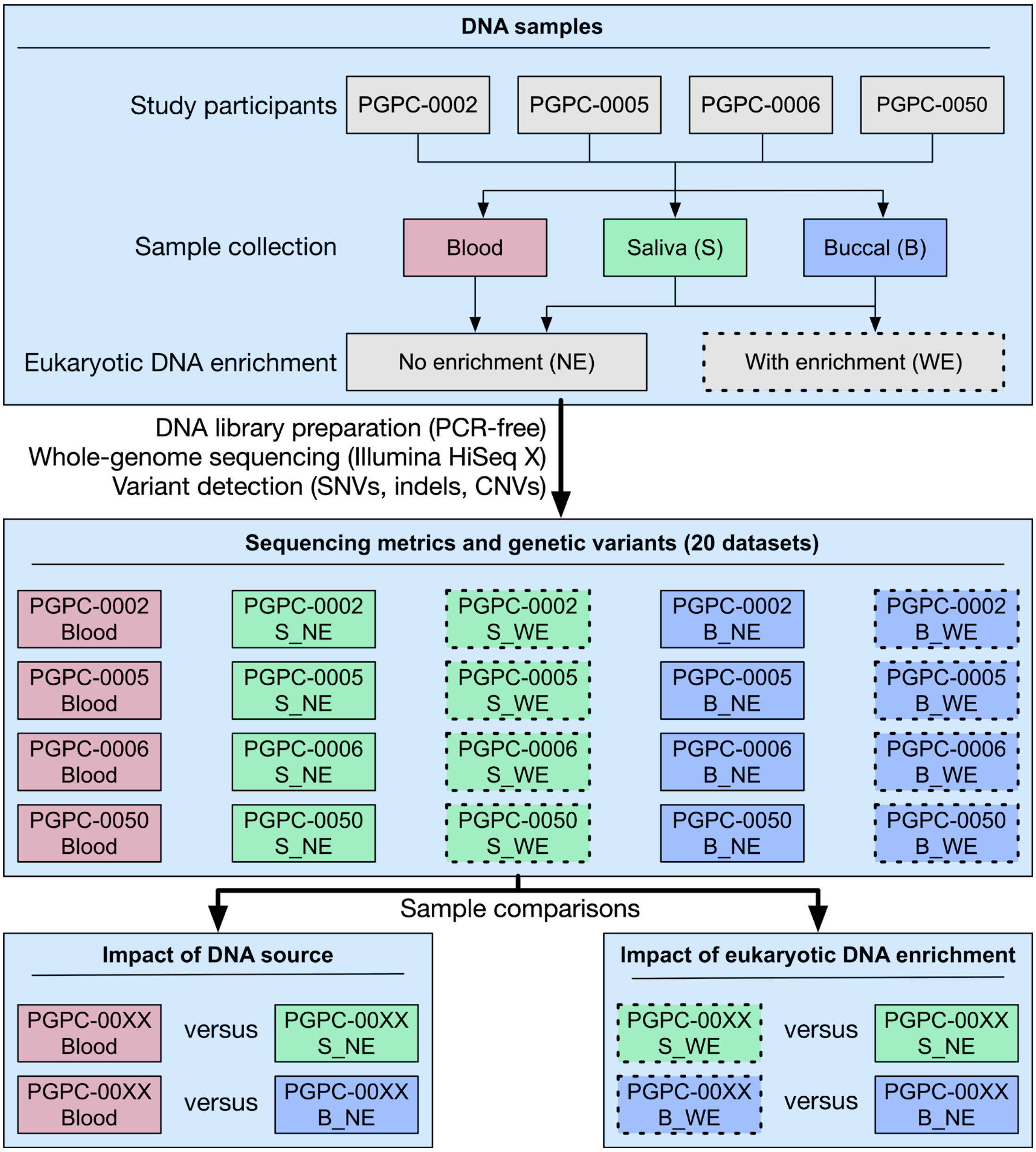

Figure 1 Study design. From each of four individuals, three sources of DNA were collected (blood, saliva and buccal). Five DNA libraries were prepared per individual—blood, saliva without eukaryotic DNA enrichment, saliva with enrichment, buccal without enrichment and buccal with enrichment. Whole-genome sequencing and genetic variant detection were performed for the 20 DNA libraries, which were compared with one another to determine the impact of DNA source and eukaryotic DNA enrichment on sequencing data quality and variant detection. B_NE, non-enriched buccal; B_WE, enriched buccal; S_NE, non-enriched saliva; S_WE, enriched saliva.

detection. As a baseline for variant-detection concordance when DNA library preparation and sequencing were performed twice for the same individual and DNA source, we used two replicates from a blood-derived HuRef sample. Concordance between blood samples and non-enriched saliva or buccal samples was similar to the baseline concordance for both SNVs and indels (table 1 and online supplementary file 3). (The HuRef blood-derived DNA replicates were sequenced nearly 3 years apart, so batch effects may explain why they did not exhibit greater concordance with each other than observed between different DNA sources.) Except for novel SNVs, concordance was similar when comparisons were restricted to coding exons, all exons, introns or intergenic regions (online supplementary tables 8-9). Compared with exons, concordance was lower in introns and intergenic regions, where increased repetitive and low-complexity elements complicate variant detection. To evaluate the accuracy of discordant variants, we used Integrative Genomics Viewer (IGV) to visually inspect read alignments for 100 SNVs and 100 indels that were detected in a blood sample but not in the corresponding non-enriched saliva or buccal sample or vice versa (online supplementary file 4). A variant was deemed false if it had little supporting evidence, if many supporting reads had poor mapping quality or were soft clipped, or if reads from one strand predominated (online supplementary figure 4). We observed no statistically significant difference in accuracy (χ^2 test) between variants detected only in blood samples and variants detected only in non-enriched saliva or buccal samples (online supplementary table 10).

Although not statistically significant, several non-enriched saliva and buccal samples had substantially more novel SNVs

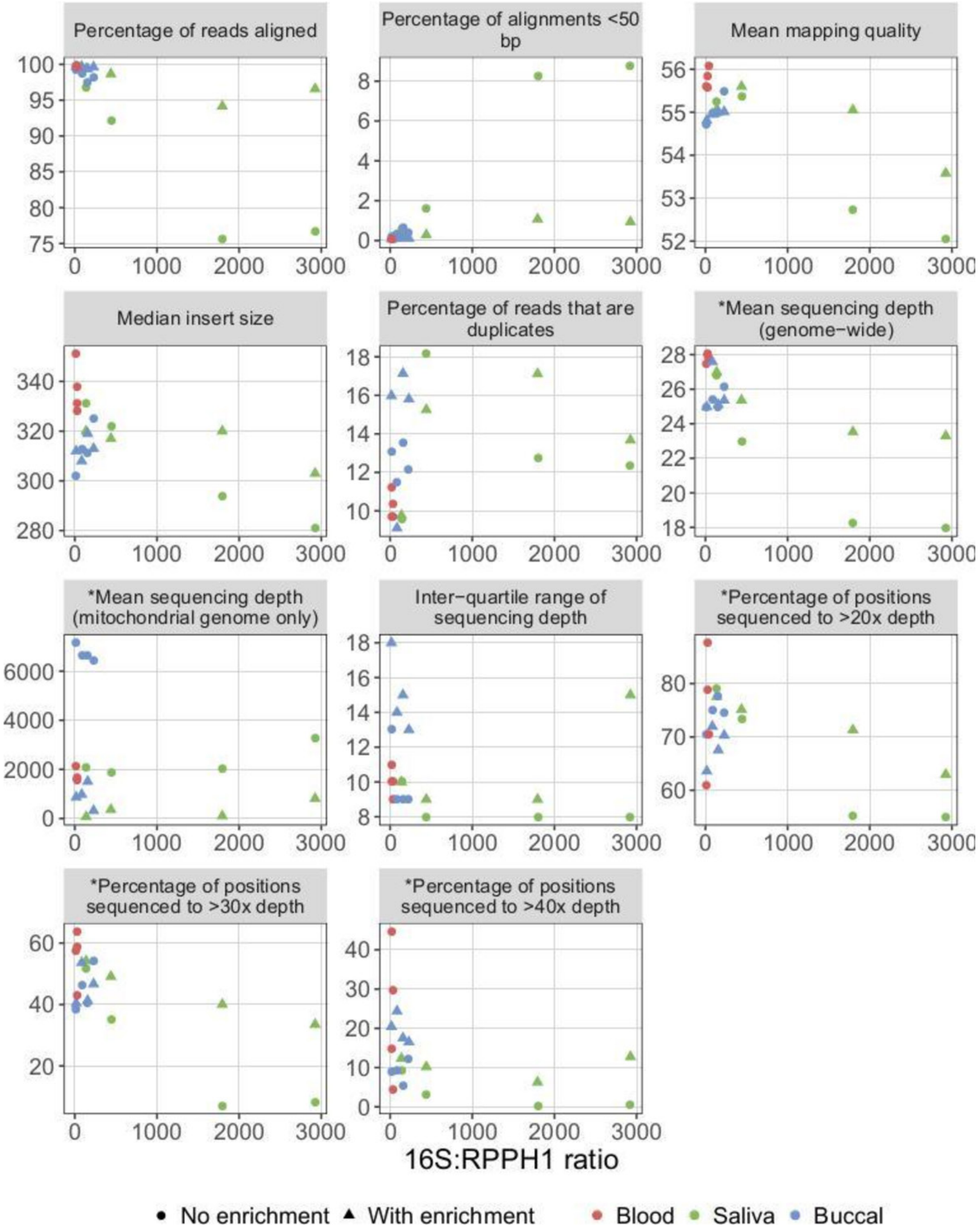

Figure 2 Relationship between bacterial DNA concentration and sequencing metrics. Higher 16S:RPPH1 ratios indicate higher bacterial DNA concentrations. Metrics prefixed with an asterisk were corrected for the total number of reads in a given sample. For saliva and buccal samples, the same sample is shown for sequencing data generated either with or without prior enrichment for eukaryotic DNA. For example, when the saliva sample with 16S:RPPH1 ratio ~2900 (online supplementary figure 1) was sequenced without first performing eukaryotic DNA enrichment, approximately 77% of reads aligned (top-left scatterplot), versus 97% when eukaryotic DNA enrichment was performed. Higher values for the inter-quartile range of sequencing depth indicate lower read-depth uniformity.

than the corresponding blood sample (table 1 and online supplementary table 6). The majority of discordant novel SNVs were false (online supplementary table 10), suggesting that some saliva and buccal samples had higher false positive rates (FPRs) for SNVs. Since known SNVs outnumbered novel SNVs by approximately 100:1 (online supplementary table 6), this difference in FPR is negligible for SNVs as a whole. However, when identifying genetic associations with disease, rare variants (eg,<1% population frequency) are often of interest. As only a small percentage of variants detected in an individual are rare, the increased FPR for novel SNVs in some saliva and buccal samples is more consequential for rare variants. Coding exons exhibited the largest differences in the number of novel SNVs detected (online supplementary table 9); aggregating over the individuals, 157 novel coding SNVs were detected in blood samples, 218 in non-enriched buccal samples and 776 in non-enriched saliva samples. We detected substantially more novel coding SNVs in non-enriched samples with high bacterial concentrations than in

Table 1 SNV- and indel-detection concordance between blood samples and non-enriched saliva or buccal samples and between enriched saliva or buccal samples and the corresponding non-enriched samples, for filtered variants detected anywhere in the genome.

		Concordant	Unique to sample type 1	Unique to sample type 2	Concordant	Unique to sample type 1	Unique to sample type 2
Sample type 1	**Sample type 2**	**Known**			**Novel**		
SNVs							
HuRef blood 1	HuRef blood 2	94.8	3.6	1.6	52.4	39.5	8.1
Blood	Non-enriched saliva	96.4	1.7	1.8	56.6	14.0	29.4
Blood	Non-enriched buccal	96.1	2.0	1.9	49.7	14.9	35.4
Enriched saliva	Non-enriched saliva	96.8	1.6	1.6	56.7	17.0	26.4
Enriched buccal	Non-enriched buccal	96.1	1.8	2.1	48.8	16.3	34.9
Indels							
HuRef blood 1	HuRef blood 2	87.4	5.9	6.7	65.4	14.2	20.3
Blood	Non-enriched saliva	87.0	5.9	7.1	63.4	16.6	20.0
Blood	Non-enriched buccal	86.4	6.4	7.2	63.4	16.2	20.3
Enriched saliva	Non-enriched saliva	87.1	6.2	6.7	63.3	18.9	17.7
Enriched buccal	Non-enriched buccal	86.4	6.6	7.0	64.5	18.0	17.4

Concordances are shown for known variants (those present in gnomAD) and novel variants. Numbers represent the percentage of variants in that category; for instance, of all known SNVs detected in either the non-enriched or the enriched buccal samples from a given individual, 96.1% were detected in both non-enriched and enriched, 1.8% were detected only in enriched, and 2.1% were detected only in non-enriched. HuRef blood 1 and HuRef blood 2 refer to replicates sequenced from the same blood-derived DNA sample and represent a baseline level of concordance; all other values were aggregated across the four study participants. For individual-specific data, see online supplementary file 3.

SNV, single nucleotide variant.

the corresponding enriched samples or in non-enriched samples with low bacterial concentrations (figure 3A). We visualised alignments for 15 novel coding SNVs detected in each individual's non-enriched saliva sample but not the corresponding blood sample, and nearly all appeared to be false variants caused by the alignment of short segments of bacteria-derived reads (figure 3B and online supplementary file 4).

We also examined the concordance of SNV and indel detection between the matched enriched and non-enriched saliva samples, and likewise for buccal samples. For both DNA sources, concordance for both SNVs and indels was similar to that of the HuRef blood replicates (table 1). Visual inspection of read alignments revealed no statistically significant difference in accuracy (χ^2 test) between variants detected only in enriched samples and those detected only in non-enriched samples (online supplementary table 10). In aggregate, substantially more novel SNVs were detected only in non-enriched samples than only in enriched samples (table 1), mirroring the comparison between blood samples and non-enriched saliva or buccal samples.

To assess sensitivity for clinically relevant variants in the four study participants, we examined 127 SNVs and 15 indels that were previously determined to be of potential clinical interest.[24] Every SNV except one was detected in all five sample types (online supplementary file 5). Eleven of the 15 indels were

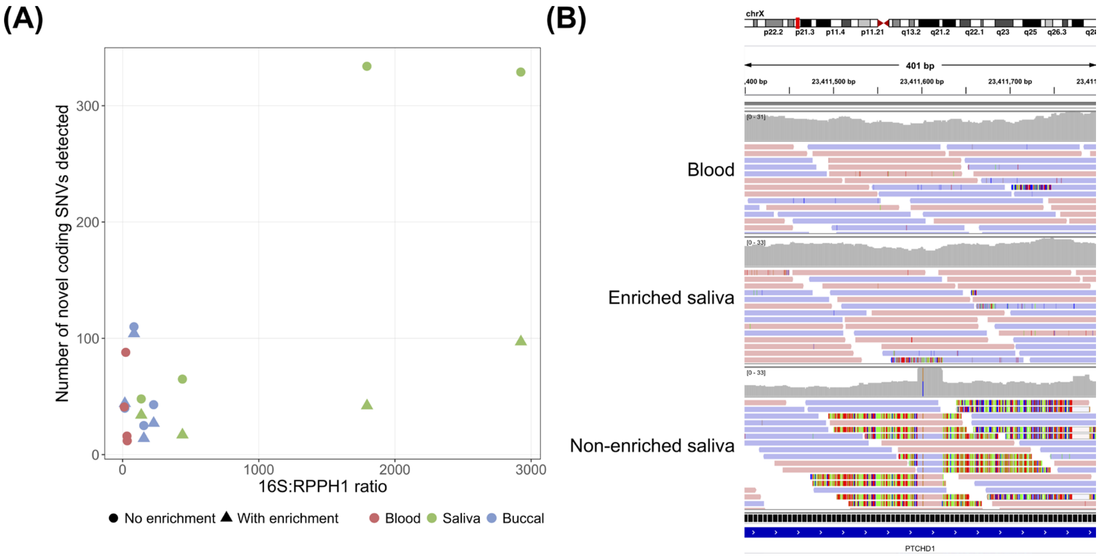

Figure 3 Bacterial contamination and the detection of false single nucleotide variants (SNVs). (A) Relationship between bacterial DNA concentration and the number of novel coding SNVs detected in each sample. For further details, see figure 2. (B) Integrative Genomics Viewer read pile-up showing a false SNV in an exon of *PTCHD1* detected in the non-enriched saliva sample from individual PGPC-0050, but not in the enriched saliva sample or blood sample from the same individual. The false SNV was detected because many short segments of bacterial reads containing a sequence difference relative to the human reference genome aligned to this region. A BLAST search suggested that the aligned bacterial reads were derived from the genome of *Fusobacterium periodonticum* (99% query cover, 97% identity), a bacterium known to be found in the human oral cavity.[45]

Table 2 Summary of the impact of DNA source and eukaryotic DNA enrichment on the accuracy of genetic variant detection from whole-genome sequencing data.

Variant type	Sensitivity	False positive rate
Blood versus non-enriched saliva or buccal		
SNVs	Little or no difference	Blood
Indels	Little or no difference	Little or no difference
CNVs (deletions)	Blood	Little or no difference
CNVs (duplications)	Blood	Blood
Enriched versus non-enriched saliva or buccal		
SNVs	Little or no difference	Enriched
Indels	Little or no difference	Little or no difference
CNVs (deletions)	Non-enriched	Little or no difference
CNVs (duplications)	Non-enriched	Non-enriched

For each comparison, the better sample type (ie, the one having higher sensitivity or a lower false positive rate) is indicated. Blood and enriched saliva and buccal samples tended to have lower false positive rates for SNVs than non-enriched saliva and buccal samples, but the magnitude of the differences were small except when considering rare SNVs (see text) and exhibited variability across individuals.
CNV, copy number variant; SNV, single nucleotide variant.

detected in every sample type; the remaining four were each missed in a single sample type (one in each type).

Finally, we assessed the impact of the differences in mitochondrial read depth among the sample types (figure 2 and online supplementary table 3) on SNV and indel detection in the mitochondrial genome. The enriched saliva samples from PGPC-0002 and PGPC-0005, which had by far the lowest mitochondrial read depths, contained two clusters of apparent SNVs, each nearly identical in the two samples, that were absent from the blood and non-enriched saliva samples from the same individuals and from the enriched saliva samples from PGPC-0006 and PGPC-0050 (online supplementary figure 5). Reads supporting these SNVs were found in all 20 samples, but comprised a much greater proportion of the reads mapping to those positions in the enriched saliva samples from PGPC-0002 and PGPC-0005 (online supplementary table 11). Reads containing these SNVs are likely derived from nuclear mitochondrial insertions,[30 31] which would explain why the numbers of reads supporting the alternate alleles were similar across samples regardless of mitochondrial read depth. The reduced mitochondrial read depth in the enriched samples may also affect heteroplasmy detection: fractions could be skewed, and low-level heteroplasmy missed altogether.

Overall, DNA source and eukaryotic DNA enrichment had a minor impact on the detection of small variants. Differences included the higher FPR for novel (especially coding) SNVs in some non-enriched saliva and buccal samples and the false mitochondrial SNVs detected in enriched samples (table 2).

Impact of DNA source and eukaryotic DNA enrichment on CNV detection

CNVs were detected using our validated workflow[28] involving the read depth-based algorithms ERDS[26] and CNVnator.[27] The number of CNVs detected differed with sample type (online supplementary table 12); in particular, the number of common CNVs (those with >1% population frequency[32]) detected in blood was typically higher than in the other sample types and lower in the enriched buccal samples.

To compare CNV detection in blood samples with that in non-enriched saliva and buccal samples, we enumerated CNVs detected concordantly or discordantly between the blood sample and the non-enriched saliva or buccal sample from the same individual. This was done for both common (table 3) and rare (online supplementary table 13) CNVs. We visually inspected alignments using IGV[28] to assess the accuracy of all discordant rare CNVs and a subset of discordant common CNVs (online supplementary file 6). Compared with the non-enriched saliva and buccal samples, CNV detection was more sensitive in blood-derived DNA, with the magnitude of the effect dependent on CNV size and type. Among the four individuals, we detected 463 common deletions between 1 and 5 kb in both blood and non-enriched saliva, 244 only in blood and 117 only in non-enriched saliva (table 3), giving a ratio of (463+244)/(463+117)=1.22 (ie, blood was 22% more sensitive than non-enriched saliva). Similarly, 25% more deletions between 1 and 5 kb were detected in blood samples than in non-enriched buccal samples. Because nearly all discordant deletions were deemed correct by visual confirmation (table 3), these disparities in the number of detected deletions constitute real sensitivity differences. For deletions between 5 and 10 kb, sensitivity in blood samples was 21% and 10% higher than in non-enriched saliva or buccal samples, respectively. Little difference was observed for deletions >10 kb. The advantage of blood samples over non-enriched saliva or buccal samples was more modest for duplications: sensitivity was 14% and 4% greater for common duplications between 1 and 5 kb, 19% and 7% greater for those between 5 and 10 kb, and nearly identical for those >10 kb. Variations among individuals generally resulted in there being no statistically significant differences in the number of deletions detected among blood, non-enriched saliva and non-enriched buccal samples (Friedman repeated-measures test; online supplementary table 14); the aggregate differences described above appear to be driven by high bacterial content in certain samples, particularly saliva (online supplementary figure 1 and online supplementary file 3).

With respect to FPRs, little difference was observed between blood samples and non-enriched saliva or buccal samples for deletions. FPRs for duplications were higher overall than for deletions, reflecting the greater difficulty of duplication detection, but were higher in non-enriched saliva and buccal samples. In particular, all rare duplications detected in non-enriched saliva or buccal samples but not in blood samples were false (online supplementary table 13).

To investigate the effect of eukaryotic DNA enrichment on CNV detection, we enumerated CNVs detected concordantly or discordantly in the enriched and non-enriched saliva samples from a given individual (and likewise for buccal). Sensitivity for both deletions and duplications was generally better in the non-enriched than in the enriched samples, particularly for buccal, for which the effect was statistically significant (Wilcoxon signed-rank test; table 3 and online supplementary table 14). Visual inspection of alignments revealed that non-uniform read depth likely explained some of the deletions missed in the enriched samples (online supplementary figure 6A), whereas others were difficult to explain (online supplementary figure 6B). For buccal samples, the FPR for large deletions was higher in enriched samples than in non-enriched (table 3 and online supplementary table 13), likely due to less uniform read depth—a trend also evident (but less pronounced) in saliva samples (online supplementary figure 6C). Poor read-depth uniformity can cause the detection of false CNVs when using PCR-based DNA library preparation,[28] and methylation-based eukaryotic DNA enrichment appeared to produce an analogous effect. FPRs for duplications were higher in enriched than in non-enriched samples, again likely due to lower read-depth uniformity (online

Table 3 Concordance between blood samples and non-enriched saliva or buccal samples and between enriched saliva or buccal samples and the corresponding non-enriched samples, for common CNVs (those with >1% frequency in MSSNG parents[32]).

		Concordant	Unique to sample type 1	Unique to sample type 2	Concordant	Unique to sample type 1	Unique to sample type 2	Concordant	Unique to sample type 1	Unique to sample type 2
Sample type 1	**Sample type 2**		**[1 kb, 5 kb)**			**[5 kb,10 kb)**			**[10 kb,…)**	
Deletions										
HuRef blood 1	HuRef blood 2	127	41 (3/3)	35 (2/3)	66	7 (3/3)	2 (2/2)	31	0 (0/0)	1 (1/1)
Blood	Non-enriched saliva	463	244 (32/32)	117 (23/23)	222	64 (29/29)	14 (13/13)	147	15 (12/12)	4 (3/4)
Blood	Non-enriched buccal	460	247 (35/35)	107 (19/19)	248	38 (23/23)	13 (10/10)	149	13 (11/11)	9 (7/9)
Enriched saliva	Non-enriched saliva	359	100 (18/18)	220 (22/23)	190	37 (21/21)	47 (19/19)	123	7 (7/7)	28 (8/9)
Enriched buccal	Non-enriched buccal	209	18 (14/14)	360 (36/37)	104	7 (6/7)	154 (34/34)	71	12 (4/9)	88 (23/24)
Duplications										
HuRef blood 1	HuRef blood 2	28	2 (1/2)	10 (1/3)	17	3 (2/3)	2 (1/2)	32	2 (0/2)	3 (0/3)
Blood	Non-enriched saliva	107	34 (10/21)	17 (6/15)	49	13 (6/13)	3 (1/3)	150	9 (1/8)	14 (0/12)
Blood	Non-enriched buccal	105	36 (7/21)	31 (5/22)	48	14 (7/14)	10 (2/10)	146	13 (0/10)	11 (1/10)
Enriched saliva	Non-enriched saliva	85	12 (5/10)	39 (7/19)	33	6 (3/6)	18 (7/12)	123	27 (1/17)	42 (4/22)
Enriched buccal	Non-enriched buccal	49	3 (0/3)	84 (12/36)	22	1 (1/1)	40 (12/27)	110	33 (0/22)	46 (5/19)

The 'concordant' columns contain the number of CNVs detected in both sample type 1 and sample type 2. The 'unique to sample type 1' columns contain the total number of CNVs detected in sample type 1 but not sample type 2, followed by an expression of the form X/Y, where X is the number of CNVs verified as correct by visual inspection and Y is the total number inspected (and analogously for the 'unique to sample type 2' columns). For example, 209 common deletions between 1 and 5 kb were detected in both the enriched buccal sample and the non-enriched buccal sample in the same individual, while 18 were detected only in the enriched buccal sample and 360 were detected only in the non-enriched buccal sample. Of the 37 deletions detected only in non-enriched buccal samples that were checked by visual confirmation, 36 were classified as true. HuRef blood 1 and HuRef blood 2 refer to replicates sequenced from the same blood-derived DNA sample and represent a baseline level of concordance. All other counts were aggregated across the four study participants. For individual-specific data, see online supplementary file 3.
CNV, copy number variant.

supplementary figure 6D); in particular, none of the rare duplications unique to the enriched samples appeared correct (online supplementary table 13).

To confirm that these differences in CNV-detection accuracy were not specific to our ERDS and CNVnator-based workflow, we detected CNVs using an alternative workflow based on Canvas.[33] Specifically, we determined the fraction of CNVs detected by Canvas in a given individual and sample type that were also detected by our standard CNV-detection workflow in the blood sample from the same individual. Blood samples were used for comparison because blood was the most accurate sample type for our standard workflow. The two approaches generally yielded consistent results: CNV detection was more sensitive in non-enriched saliva or buccal samples compared with enriched, and sensitivity in blood samples was higher than in non-enriched buccal samples (although blood and non-enriched saliva samples had similar sensitivity with Canvas) (online supplementary table 15).

Overall, DNA source and eukaryotic DNA enrichment had a more substantial impact on the read depth-based detection of CNVs than they did on small variants, with higher accuracy in blood samples than in non-enriched saliva or buccal samples and higher accuracy in non-enriched than in enriched samples (table 2).

Impact of DNA source and eukaryotic DNA enrichment on structural variation (SV) detection

In this study, we concentrated on SNVs, indels and CNVs, as there exist fully established workflows for their detection.[25 28 34] As a preliminary investigation into the effect of sample type on SV detection, we employed Manta,[35] which uses anomalously mapped paired-end reads and soft-clipped reads to detect SVs. Specifically, we enumerated SVs of each type (deletions, duplications, inversions, insertions and breakends) detected by Manta in each sample as a crude measure of sensitivity. For comparison with the read-depth results, deletion and duplication counts were stratified by size. In general, more deletions and duplications were detected in blood samples than in the other sample types, although the magnitudes of the differences were generally small and varied by size (online supplementary table 16). The small differences in apparent sensitivity among sample types suggest that methods based on anomalously mapped paired-end reads and soft-clipped reads may be able to partially compensate for the reduced sensitivity of deletion and duplication detection observed in the non-blood sample types when using read depth-based approaches. For other SV types, more variants were detected in the blood sample than in any of the other sample types in 3/4 individuals for inversions, 0/4 for insertions and 2/4 for breakends. Once reliable, validated workflows for SV detection have been established, we will more thoroughly investigate the effect of DNA source and eukaryotic DNA enrichment using the same methodology employed for SNVs, indels and CNVs.

DISCUSSION

In the design and implementation of our own WGS studies[32 36 37] and in running a service-based sequencing centre, questions often arise about whether saliva- or buccal-collection kits yield DNA sufficient for comprehensive WGS and genetic variant detection, and how these data compare with those from the current gold standard (blood-derived DNA). To investigate the impact of

DNA source for researchers and clinicians, five sample types—blood, saliva with or without methylation-based eukaryotic DNA enrichment, and buccal swabs with or without enrichment—were sequenced from each of four individuals. Blood consistently gave the best sequencing metrics, and although enrichment of saliva or buccal samples decreased the percentage of unmapped reads and short, spurious alignments, it also reduced read-depth uniformity and mitochondrial read depth. Consistent with Wall *et al*,[11] DNA source had little effect on the accuracy of SNV detection, although we found that the FPR for rare SNVs was higher in some non-enriched saliva and buccal samples. However, DNA source appeared to affect the accuracy of read depth-based CNV detection—sensitivity for deletions and duplications was higher in blood samples than in non-enriched saliva or buccal samples, and the FPR for duplications was lower in blood samples. Eukaryotic DNA enrichment hampered read depth-based CNV discovery, with non-enriched samples giving better sensitivity for deletions and duplications and a lower FPR for duplications. The reduced accuracy of read depth-based CNV detection in enriched samples was likely due to lower read-depth uniformity, which may result from non-uniform methylation causing some genomic regions to be captured more efficiently than others.[38 39] In this study, the sequencing data were subsampled to eliminate read depth as a confounding factor when evaluating variant-detection accuracy. Had this step been omitted, we might have observed larger differences among sample types, especially for samples with high bacterial concentrations. This possibility is supported by our previous study, in which sensitivity for detecting deletions<10 kb decreased when the mean read depth was less than ~30×.[28]

Besides variant-detection accuracy, other considerations may be important when choosing a DNA source. Blood is collected by a professional phlebotomist, leaving little risk of improper collection. For saliva or buccal samples, participants may provide too little material, especially when self-collected, or may ignore instructions to refrain from eating. However, blood can be difficult to collect from individuals who fear needles and from children with behavioural difficulties or sensitivity to touch or pressure. Saliva and buccal samples are more stable than blood samples, can be collected in the participant's home (for research purposes) and can be shipped more easily. In clinical diagnostics, additional factors may influence the choice of DNA source. For instance, certain neurodevelopmental and neurological disorders have causative variants specific to, or more evident in, certain sample types, such as ectodermal-derived tissues (which include buccal cells).[40–42] When detecting somatic mutations in patients with leukaemia, blood cannot be used as a matched normal sample. For mitochondrial variants, heteroplasmy can vary across tissue types.[43] If saliva or buccal samples are preferred given these considerations, then we recommend against methylation-based eukaryotic DNA enrichment, as the advantages of enrichment appear negligible and are outweighed by the drawbacks noted above. By aligning against the human reference genome, most bacterial reads are removed automatically. Increases in read depth with enrichment were modest; the same increase could be achieved via additional sequencing—an option that will become even more appealing as sequencing costs continue to decline.

Unless saliva or buccal samples are preferred for reasons such as those outlined above, we recommend using DNA derived from blood samples for WGS, as it equalled or surpassed saliva and buccal samples (although often only slightly) for all comparisons performed in this study. As more WGS data sets are generated, the ability to accurately detect genetic variants of all types will be critically important for population genetics studies, disease studies and clinical diagnostics. Large-scale meta-analyses will become increasingly valuable; however, a significant challenge is data heterogeneity, which can originate from differences in DNA library preparation, sequencing platform, read depth, etc. Although methods exist for addressing heterogeneity,[44] it is undoubtedly beneficial to remove its sources in advance. Given that differing DNA sources add heterogeneity, and that whole-blood samples appear to be better than saliva and buccal samples for WGS, continued use of blood as the first-line tissue source would facilitate accurate, large-scale comparative analyses of WGS data.

Acknowledgements We thank Janet Buchanan and Richard Wintle for critical feedback on the manuscript; Jo-Anne Herbrick, Jennifer Howe and Barbara Kellam for helpful discussions and Zhuozhi Wang for helping deposit the sequencing data to NCBI.

Contributors BT, SW and SWS designed the study. SW and MJZ managed ethics approval and sample collection. BT, SW, SAH, WWLS, SP, CLP, EJH, LJS, CN, AR and CRM analysed the data. BT and SWS supervised the study and wrote the manuscript. All authors have read and approved the final manuscript.

Funding BT is funded by the Canadian Institutes of Health Research (CIHR) Banting Postdoctoral Fellowship. SWS is funded by the GlaxoSmithKline-CIHR Chair in Genome Sciences at the University of Toronto and The Hospital for Sick Children. Technology development funds for this research were used from grants from Genome Canada, the University of Toronto McLaughlin Centre, and The Hospital for Sick Children Foundation. The sequencing and informatics infrastructure was supported by the Canada Foundation for Innovation.

Competing interests None declared.

Ethics approval This study was approved by the Research Ethics Board (REB) at The Hospital for Sick Children (REB no. 1000053640).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

ORCID iDs
Brett Trost http://orcid.org/0000-0003-4863-7273
Edward J Higginbotham http://orcid.org/0000-0003-1881-9308

REFERENCES

1 Kan YW, Dozy AM, Trecartin R, Todd D. Identification of a nondeletion defect in alpha-thalassemia. *N Engl J Med* 1977;297:1081–4.
2 Sun F, Reichenberger EJ. Saliva as a source of genomic DNA for genetic studies: review of current methods and applications. *Oral Health Dent Manag* 2014;13:217–22.
3 Meghnani V, Mohammed N, Giauque C, Nahire R, David T. Performance characterization and validation of saliva as an alternative specimen source for detecting hereditary breast cancer mutations by next generation sequencing. *Int J Genomics* 2016;2016:1.
4 Perkins BA, Caskey CT, Brar P, Dec E, Karow DS, Kahn AM, Hou Y-CC, Shah N, Boeldt D, Coughlin E, Hands G, Lavrenko V, Yu J, Procko A, Appis J, Dale AM, Guo L, Jönsson TJ, Wittmann BM, Bartha I, Ramakrishnan S, Bernal A, Brewer JB, Brewerton S, Biggs WH, Turpaz Y, Venter JC. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc Natl Acad Sci U S A* 2018;115:3686–91.
5 Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* 2018;15:353–65.
6 Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* 2018;19:110–24.
7 Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, Lee MK, Zaidi A, Mattern BC, Liebowitz C, Pearson L, González T, Leslie EJ, Carlson JC, Orlova E, Suetens P, Vandermeulen D, Feingold E, Marazita ML, Shaffer JR, Wysocka J, Shriver MD, Weinberg SM. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat Genet* 2018;50:414–23.
8 Bahlo M, Stankovich J, Danoy P, Hickey PF, Taylor BV, Browning SR, Brown MA, Rubio JP. Saliva-derived DNA performs well in large-scale, high-density single-nucleotide polymorphism microarray studies. *Cancer Epidemiol Biomark Prev* 2010;19:794–8.
9 Yokoyama JS, Erdman CA, Hamilton SP. Array-based whole-genome survey of dog saliva DNA yields high quality SNP data. *PLoS One* 2010;5:e10809.

10 Abraham JE, Maranian MJ, Spiteri I, Russell R, Ingle S, Luccarini C, Earl HM, Pharoah PPD, Dunning AM, Caldas C. Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics* 2012;5.

11 Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* 2014;24:1734–9.

12 Gudiseva HV, Hansen M, Gutierrez L, Collins DW, He J, Verkuil LD, Danford ID, Sagaser A, Bowman AS, Salowe R, Sankar PS, Miller-Ellis E, Lehman A, O'Brien JM. Saliva DNA quality and genotyping efficiency in a predominantly elderly population. *BMC Med Genomics* 2016;9.

13 Reiner J, Karger L, Cohen N, Mehta L, Edelmann L, Scott SA. Chromosomal microarray detection of constitutional copy number variation using saliva DNA. *J Mol Diagn* 2017;19:397–403.

14 Bruinsma FJ, Joo JE, Wong EM, Giles GG, Southey MC. The utility of DNA extracted from saliva for genome-wide molecular research platforms. *BMC Res Notes* 2018;11:8.

15 Woo JG, Sun G, Haverbusch M, Indugula S, Martin LJ, Broderick JP, Deka R, Woo D. Quality assessment of buccal versus blood genomic DNA using the Affymetrix 500 K GeneChip. *BMC Genet* 2007;8:79.

16 Rincon G, Tengvall K, Belanger JM, Lagoutte L, Medrano JF, André C, Thomas A, Lawley CT, Hansen MST, Lindblad-Toh K, Oberbauer AM. Comparison of buccal and blood-derived canine DNA, either native or whole genome amplified, for array-based genome-wide association studies. *BMC Res Notes* 2011;4.

17 Feigelson HS, Rodriguez C, Welch R, Hutchinson A, Shao W, Jacobs K, Diver WR, Calle EE, Thun MJ, Hunter DJ, Thomas G, Chanock SJ. Successful genome-wide scan in paired blood and buccal samples. *Cancer Epidemiol Biomark Prev* 2007;16:1023–5.

18 Loomis SJ, Olson LM, Pasquale LR, Wiggs J, Mirel D, Crenshaw A, Parkin M, Rahhal B, Tetreault S, Kraft P, Tworoger SS, Haines JL, Kang JH. Feasibility of high-throughput genome-wide genotyping using DNA from stored buccal cell samples. *Biomark Insights* 2010;5:BMI.S5062–55.

19 Erickson SW, MacLeod SL, Hobbs CA. Cheek swabs, SNP chips, and CNVs: assessing the quality of copy number variant calls generated with subject-collected mail-in buccal brush DNA samples on a high-density genotyping microarray. *BMC Med Genet* 2012;13.

20 Fabre A, Thomas E, Baulande S, Sohier E, Hoang L, Soularue P, Ragusa S, Clavel-Chapelon F, Cox DG. Is saliva a good alternative to blood for high density genotyping studies: SNP and CNV comparisons. *J Biotech Biomat* 2011;1.

21 Hu Y, Ehli EA, Nelson K, Bohlen K, Lynch C, Huizenga P, Kittelsrud J, Soundy TJ, Davies GE. Genotyping performance between saliva and blood-derived genomic DNAs on the DMET array: a comparison. *PLoS One* 2012;7:e33968.

22 Pennell CE, Vadillo-Ortega F, Olson DM, Ha E-H, Williams S, Frayling TM, Dolan S, Katz M, Merialdi M, Menon R. Preterm Birth Genome Project (PGP) – validation of resources for preterm birth genome-wide studies. *J Perinat Med* 2013;41:45–9.

23 Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA, Pradhan S. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One* 2013;8:e76096.

24 Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheimer N, Yuen RKC, Trost B, Paton TA, Pereira SL, Herbrick J-A, Wintle RF, Merico D, Howe J, MacDonald JR, Lu C, Nalpathamkalam T, Sung WWL, Wang Z, Patel RV, Pellecchia G, Wei J, Strug LJ, Bell S, Kellam B, Mahtani MM, Bassett AS, Bombard Y, Weksberg R, Shuman C, Cohn RD, Stavropoulos DJ, Bowdin S, Hildebrandt MR, Wei W, Romm A, Pasceri P, Ellis J, Ray P, Meyn MS, Monfared N, Hosseini SM, Joseph-George AM, Keeley FW, Cook RA, Fiume M, Lee HC, Marshall CR, Davies J, Hazell A, Buchanan JA, Szego MJ, Scherer SW. The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. *Can Med Assoc J* 2018;190:E126–E136.

25 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.

26 Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, Ruzzo EK, Gumbs C, Singh A, Feng S, Shianna KV, Goldstein DB. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* 2012;91:408–21.

27 Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;21:974–84.

28 Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, Pereira SL, Whitney J, Chan AJS, Pellecchia G, Reuter MS, Lok S, Yuen RKC, Marshall CR, Merico D, Scherer SW. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet* 2018;102:142–55.

29 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–91.

30 Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (NuMts) in sequenced nuclear genomes. *PLoS Genet* 2010;6:e1000834.

31 Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res* 2014;42:12640–9.

32 Yuen RKC, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G, Buchanan JA, Walker S, Marshall CR, Uddin M, Zarrei M, Deneault E, D'Abate L, Chan AJS, Koyanagi S, Paton T, Pereira SL, Hoang N, Engchuan W, Higginbotham EJ, Ho K, Lamoureux S, Li W, MacDonald JR, Nalpathamkalam T, Sung WWL, Tsoi FJ, Wei J, Xu L, Tasse A-M, Kirby E, Van Etten W, Twigger S, Roberts W, Drmic I, Jilderda S, Modi BM, Kellam B, Szego M, Cytrynbaum C, Weksberg R, Zwaigenbaum L, Woodbury-Smith M, Brian J, Senman L, Iaboni A, Doyle-Thomas K, Thompson A, Chrysler C, Leef J, Savion-Lemieux T, Smith IM, Liu X, Nicolson R, Seifer V, Fedele A, Cook EH, Dager S, Estes A, Gallagher L, Malow BA, Parr JR, Spence SJ, Vorstman J, Frey BJ, Robinson JT, Strug LJ, Fernandez BA, Elsabbagh M, Carter MT, Hallmayer J, Knoppers BM, Anagnostou E, Szatmari P, Ring RH, Glazer D, Pletcher MT, Scherer SW. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* 2017;20:602–11.

33 Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* 2016;32:2375–7.

34 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma* 2013;43:1–33.

35 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;32:1220–2.

36 Yuen RKC, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, Chrysler C, Nalpathamkalam T, Pellecchia G, Liu Y, Gazzellone MJ, D'Abate L, Deneault E, Howe JL, Liu RSC, Thompson A, Zarrei M, Uddin M, Marshall CR, Ring RH, Zwaigenbaum L, Ray PN, Weksberg R, Carter MT, Fernandez BA, Roberts W, Szatmari P, Scherer SW. Whole-Genome sequencing of quartet families with autism spectrum disorder. *Nat Med* 2015;21:185–91.

37 Yuen RKC, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, Tong X, Sun Y, Cao D, Zhang T, Wu X, Jin X, Zhou Z, Liu X, Nalpathamkalam T, Walker S, Howe JL, Wang Z, MacDonald JR, Chan AJS, D'Abate L, Deneault E, Siu MT, Tammimies K, Uddin M, Zarrei M, Wang M, Li Y, Wang J, Wang J, Yang H, Bookman M, Bingham J, Gross SS, Loy D, Pletcher M, Marshall CR, Anagnostou E, Zwaigenbaum L, Weksberg R, Fernandez BA, Roberts W, Szatmari P, Glazer D, Frey BJ, Ring RH, Xu X, Scherer SW. Genome-wide characteristics of *de novo* mutations in autism. *NPJ Genom Med* 2016;1:160271–10.

38 Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. Large-scale structure of genomic methylation patterns. *Genome Res* 2006;16:157–63.

39 Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A. Charting a dynamic DNA methylation landscape of the human genome. *Nature* 2013;500:477–81.

40 Helgeson M, Keller-Ramey J, Knight Johnson A, Lee JA, Magner DB, Deml B, Deml J, Hu Y-Y, Li Z, Donato K, Das S, Laframboise R, Tremblay S, Krantz I, Noon S, Hoganson G, Burton J, Schaaf CP, del Gaudio D. Molecular characterization of HDAC8 deletions in individuals with atypical Cornelia de Lange syndrome. *J Hum Genet* 2018;63:349–56.

41 Huisman SA, Redeker EJW, Maas SM, Mannens MM, Hennekam RCM. High rate of mosaicism in individuals with Cornelia de Lange syndrome. *J Med Genet* 2013;50:339–44.

42 Rivière J-B, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartzentruber JA, Gripp KW, Nikkel SM, Worthylake T, Sullivan CT, Ward TR, Butler HE, Kramer NA, Albrecht B, Armour CM, Armstrong L, Caluseriu O, Cytrynbaum C, Drolet BA, Innes AM, Lauzon JL, Lin AE, Mancini GMS, Meschino WS, Reggin JD, Saggar AK, Lerman-Sagie T, Uyanik G, Weksberg R, Zirn B, Beaulieu CL, Majewski J, Bulman DE, O'Driscoll M, Shendure J, Graham JM, Boycott KM, Dobyns WB. *De novo* germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat Genet* 2012;44:934–40.

43 de Laat P, Koene S, van den Heuvel LPWJ, Rodenburg RJT, Janssen MCH, Smeitink JAM. Clinical features and heteroplasmy in blood, urine and saliva in 34 Dutch families carrying the m.3243A > G mutation. *J Inherit Metab Dis* 2012;35:1059–69.

44 Derkach A, Chiang T, Gong J, Addis L, Dobbins S, Tomlinson I, Houlston R, Pal DK, Strug LJ. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* 2014;30:2179–88.

45 Slots J, Potts TV, Mashimo PA, periodonticum F. A new species from the human oral cavity. *J Dent Res* 1983;62:960–3.