Laboratory Investigation (2014) 94, 1173–1183

Detection of somatic mutations in tumors using unaligned clonal sequencing data

Kate M Sutton[1], Laura A Crinnion[2], David Wallace[2], Sally Harrison[3], Paul Roberts[2], Christopher M Watson[2], Alexander F Markham[3], David T Bonthron[3], Philip Quirke[1] and Ian M Carr[3]

Most cancers arise and evolve as a consequence of somatic mutations. These mutations influence tumor behavior and clinical outcome. Consequently, there is considerable interest in identifying somatic variants within specific genes (such as *BRAF*, *KRAS* and *EGFR*) so that chemotherapy can be tailored to the patient's tumor genotype rather than using a generic treatment based on histological diagnosis alone. Owing to the heterogeneous nature of tumors, a somatic mutation may be present in only a subset of cells, necessitating the use of quantitative techniques to detect rare variants. The highly quantitative nature of next-generation sequencing (NGS), together with the ability to multiplex numerous samples, makes NGS an attractive choice with which to screen for somatic variants. However, the large volumes of sequence data present significant difficulties when applying NGS for the detection of somatic mutations. To alleviate this, we have developed methodologies including a set of data analysis programs, which allow the rapid screening of multiple formalin-fixed, paraffin-embedded samples for the presence of specified somatic variants using unaligned Illumina NGS data.
Laboratory Investigation (2014) **94,** 1173–1183; doi:10.1038/labinvest.2014.96; published online 28 July 2014

Deleterious sequence variants have an important part in the initiation and progression of many different tumor types. With the advent of drugs effective against specific molecular targets, there is much interest in detecting variants in specific genes (such as *BRAF*, *KRAS* and *EGFR*), so that therapy can be tailored to the patient's tumor genotype rather than relying on empirical treatments based on tumor site and histological type.[1] Clinically important variants may either be inherited through the germline or occur as spontaneous somatic mutations.[2] As germline variants are present in normal tissue, at well-defined allelic ratios, their detection by DNA screening is simple compared with the detection of somatic mutations; the latter are at best restricted to malignant or pre-malignant tissue, and at worst may occur in only a small proportion of cells within the affected tissue. Inherited germline variants can be identified using standard diagnostic screening techniques (Sanger sequencing of PCR amplicons containing the gene sequences of interest or next-generation exome sequencing,[3,4] Typically, in either case DNA will be obtained from the patient's peripheral blood. However, these approaches are not well suited to the detection of somatic mutations in heterogeneous tumor samples, where clinically important mutations may be present at a level lower than the sensitivity of PCR/Sanger or exome sequencing can detect. As only a minority of cancers are caused by germline variants, there is an urgent need to develop screening methodologies for the detection of the much more prevalent somatic mutations.

Somatic mutations typically drive carcinogenesis by one of two actions: deactivation of a protein that normally suppresses tumorigenesis; or constitutive activation of a protein such that it drives carcinogenesis.[2] Although deactivation of a protein may be caused by a large number of possible mutations across its gene, protein activation is generally caused by a specific set of sequence variants that occur at specific positions within it. For example, 80% of deactivating mutations leading to cancer occur between residues 126 and 306 in the p53 protein,[5] while 80% of activating mutations occur at residue 600 in the BRAF protein.[6] These genetic differences have implications for the detection of clinically important somatic mutations, with the entire coding sequences of some genes needing to be screened, while only specific nucleotide positions need to be analyzed in others.

[1]Section of Pathology and Tumour Biology, Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds, UK; [2]Yorkshire Regional Genetics Service, St James's University Hospital, Leeds, UK and [3]Section of Genetics, Institute of Biomedical and Clinical Sciences, School of Medicine, University of Leeds, St James's University Hospital, Leeds, UK
Correspondence: Dr IM Carr, BSc, PhD, Institute of Biomedical and Clinical Sciences, School of Medicine, University of Leeds, St James's University Hospital, Level 9 WTBB, Leeds, West Yorkshire LS9 7TF, UK.
E-mail: I.M.Carr@leeds.ac.uk

Received 19 March 2014; revised 4 June 2014; accepted 9 June 2014

As approximately one-third of individuals in developed countries are expected to develop cancer,[7] cheap and sensitive, high-throughput somatic mutation detection methodologies are required. As massively parallel sequencing technologies have matured, many diagnostic testing centers plan to move to techniques based on next-generation sequencing (NGS), as they promise high sensitivity combined with very high throughput, at low cost per sample. To achieve the required sensitivity at reasonable cost, it is necessary to enrich a sample for the desired target sequences before analysis. There are a number of approaches currently used to enrich for target sequences, including PCR amplification, molecular inversion probes, hybrid capture and in-solution hybridization capture.[8] Typically, hybridization capture reagents are used when screening megabase-sized targets, while PCR amplicon-based methods are used for comparatively small targets up to the size of several kb.

Although sequence generation has become increasingly simple, the bioinformatic analysis of these data remain a major bottleneck. To overcome this obstacle, a number of different pipelines have been developed for detecting and cataloguing sequence variants, not only for targeted somatic mutation detection, but for the technically similar goals of microbiome profiling[9,10] and HLA typing.[11,12] In each case, the analysis includes three basic steps: (i) determining the origin of each read, (ii) identifying variants within each read and (iii) aggregating the variant data to arrive at a conclusion. To determine the point of origin of a sequence within a genome, any of many sequence aligners may be used; however, when sequencing small amplicons (<500 bp) a read's point of origin may be more quickly deduced by comparing its sequence with a table of all subsequences (including known common variants) of the target amplicon(s).[13] Similarly, the presence of a sequence variant may be deduced either by comparing each read with a wild-type reference sequence[14] or by matching grouped identical reads to an index table derived from the variant reference sequence of interest.[13] Typically, to distinguish true positives, the number of reads that suggest a specific variant is counted and then compared with the total number of reads mapping to the same location; such true positives are then annotated and exported to a results file.

There are biological factors that limit the utility of software in detecting rare sequence variants. Most importantly, tumor genotyping is often performed on poor quality DNA extracted from formalin-fixed, paraffin-embedded (FFPE) samples, which is known to introduce sequence artifacts.[15–17] In addition, owing to heterogeneity of tumor subclones and admixture with normal cells, the results of analysis of a single sample may not be representative of the tumor as a whole. To maximize the reliability of a clinical test, knowledge of these biological variables will be needed, as well as optimization of the variant-calling software itself.

As far as the method used for somatic mutation detection is concerned, usability is thus likely to be a more significant issue than raw speed or detection limit. To maximize usability, we have developed AgileSMPoint and AgileSMAll, which allow the analysis to be performed on a typical desktop computer as a single operator step. These programs identify somatic sequence variants from unaligned sequence data generated from targeted amplicon libraries. As far as possible, both programs can identify and ignore sequences derived from highly homologous pseudogenes. AgileSMPoint is designed to identify sequence variants at specific positions in an amplicon, such as known activating mutations in genes like *KRAS*. AgileSMAll, in contrast, is designed to identify sequence variants at all positions within an amplicon, and so is more suited to the screening of tumor-suppressor genes such as *TP53*. We have also streamlined the production of target libraries using the method of targeted amplicon library creation, in which the Illumina adaptors and index sequences are incorporated into the amplicon during a single-step amplification process.

MATERIALS AND METHODS

Patients

Anonymized FFPE tumor samples that had previously undergone routine diagnostic somatic mutation screening in the diagnostic laboratory (Yorkshire Regional Genetics Service), using either pyrosequencing (Pyrosequencing AB, Uppsala, Sweden) or Rotor-Gene Q (Qiagen, Venlo, The Netherlands); their mutational status was consequently known. This allowed the selection of samples with a wide range of different sequence variants in the *BRAF*, *EGFR* and *KRAS* genes.

DNA Extraction

FFPE tissue was processed to produce $10 \times 5\,\mu$m sections of tissue for DNA extraction. Where necessary, the sections were macro-dissected using hematoxylin and eosin-stained sections as a guide. DNA extraction was performed using the Qiagen micro-kit (Qiagen, Manchester, UK) and resuspended in $20\,\mu$l nuclease-free water. Finally, the DNA was quantified by Quant-iT PicoGreen assay (Invitrogen, Paisley, UK) and diluted to $10\,\text{ng}/\mu\text{l}$.

Targeted Amplicon Library Creation

Library generation from genomic DNA for each of the specified targets was performed in a one-step 'touchdown' PCR. The method is summarized in Figure 1 and the PCR primers, library adaptor oligonucleotide sequences and PCR thermocyling conditions are described in Supplementary Tables S1–S3. In brief, the amplicons were amplified using PCR primers containing a 5′ tag that was not complementary to the target sequence (Supplementary Table S1). Instead, the tag in one primer was complementary to the 3′ end of an oligonucleotide that contains the universal Illumina library adaptor sequence, while the other PCR primer's tag was complementary to the 3′ end of an oligonucleotide that contained the indexed Illumina library adaptor sequence (Supplementary Table S2). Therefore, amplicons were

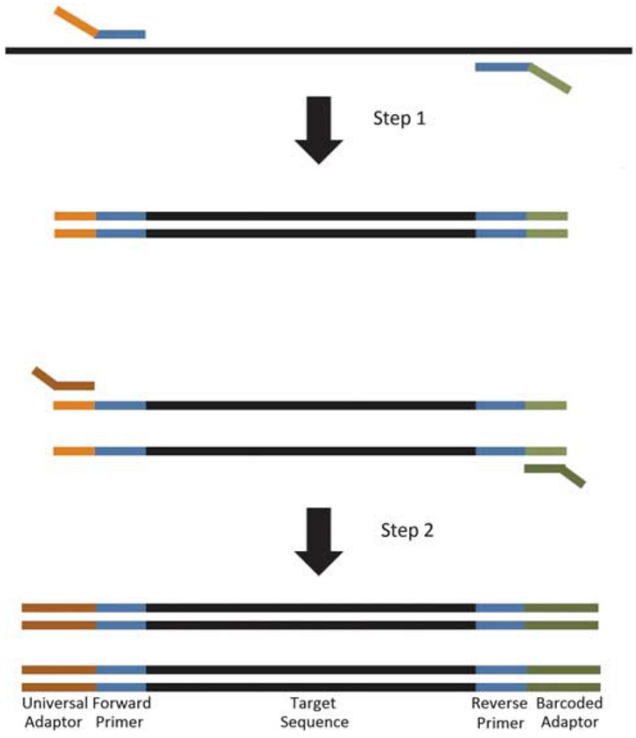

Figure 1 The amplicons were generated in a two-step process. The initial stage of amplification uses primers homologous to the target gene sequence (blue line in step 1), with a 5′ tag complementary to the 3′ end of either the Universal sequencing adaptor (pale brown line in step 1) or the bar-coded sequencing adaptor (pale green line in step 1). The resulting amplicons are subsequently amplified by primers homologous to the full-length universal sequencing adaptor (dark brown line in step 2) or the bar-coded sequencing adaptor (dark green line in step 2).

initially generated by the two target-specific primers annealing to the template DNA (Figure 1, step 1). In subsequent rounds of amplification, amplicons were also generated by the two library adaptor oligonucleotides annealing to PCR products and so incorporating the library adaptors into the amplicon (Figure 1, step 2). When multiplexing amplicons from multiple samples, each sample was identified by using an indexed adaptor with a different 6-nt index sequence for each sample. The touchdown PCR reagents and thermocycling conditions are shown in Supplementary Tables S3 and S4, respectively.

NGS of Amplicon Libraries

Each amplicon library was size selected using agarose gel electrophoresis, to remove unincorporated oligonucleotides and PCR products lacking adaptors. Then each PCR library was quantified using a Bioanalyser (Agilent Technologies, Santa Clara, CA, USA) and combined to form pools for each sample containing products at equimolar concentrations. Before 150-bp paired-end sequencing on an Illumina MiSeq (Illumina, San Diego, CA, USA), the sample pools were combined to form a single sequencing pool. The resultant sequencing reads were then de-multiplexed using the Illumina de-tagging pipeline, such that sequence data for each sample were placed in a single fastq file. Each member of a read pair was analyzed independently, whether overlapping or not.

Identification of Somatic Sequence Variants using a Standard BWA/VarScan-Based Pipeline

To create a comparison data set of somatic sequence variants to which the results generated by the software describe in this paper could be described, we aligned the sequence data to the human genome (hg19) using BWA[18] and used samtools[19] to create a pileup file. Finally, VarScan[14] was used to identify sequence variants in the pileup file. Each step was accomplished using the command line arguments shown below:

BWA (two step alignment via *.sai file)
bwa aln [index file] -f [output.sai] [data.fastq.gz]
bwa samse -f [Export.sam] -r [index file] [output.sai] [data.fastq.gz]
Samtools (pileup file creation).
samtools mpileup -f [index file] [Export.sam] > [Export.pileup]
Varscan: identification of single base changes (line A) and indels (line B)
A: java -jar VarScan.v2.3.6.jar mpileup2snp [Export.Pileup] –min-coverage 100 –min-var-freq 0.05 –min-freq-for-hom 0.95 –strand-filter 0 > ExportedVariants.vcf
B: java -jar VarScan.v2.3.6.jar mpileup2indel [Export.Pileup] –min-coverage 100 –min-var-freq 0.05 –min-freq-for-hom 0.95 –strand-filter 0 > ExportedVariants.vcf

When calling the variants, the minimum read depth was set to 100 reads and the variant allele must be present in > 5% of the reads.

Identification of Somatic Sequence Variants in Unaligned Sequence Data

Rather than detecting somatic mutations by first aligning the reads to a reference sequence and then interrogating the aligned data at each position of interest, our approach simultaneously identifies the origin of each read and maintains a running total of the base calls at positions of interest. Once the unaligned data file has been completely read, the nucleotide read depth data can then be used to assess the presence of any sequence variants. The way in which each program analyses the data are outlined below.

AgileSMPoint: identification of somatic sequence variants at mutational hot spots

Information on the sequence of each amplicon and the positions to be queried is imported into AgileSMPoint using a 'target' file, within which the information for each amplicon is present in two successive lines (Figure 2a). The first line specifies the amplicon's experimental ID, the location(s) of the nucleotide(s) of interest on a reference sequence, the

Figure 2 The format of the target files used to describe the amplicons analyzed by AgileSMPoint (**a**) and AgileSMAll (**b**).

reference allele for the nucleotide(s) of interest and the amplicon's analysis set name. A further field containing the character 'P' may be present (underlined in blue in Figure 2a). This optional field instructs AgileSMPoint to permit the presence of a SNP in the sequence flanking the positions of interest when identifying reads originating from that amplicon. The second line contains the sequence flanking the query positions of interest, which are identified by the character 'N' (underlined in red in Figure 2a). If there is a possibility that reads may also originate from homologous non-target sequences (such as pseudogenes), positions where the duplicated sequences diverge can be identified by using a lower case letter in the reference sequence (underlined in black in Figure 2a). If the sequence of a read differs at these lower case positions, the read is flagged as originating from a duplicated sequence and is excluded from subsequent analysis.

Each read is scanned for the presence of flanking sequences identical to the 18 bases preceding the first query position and the 18 bases following the last query position for each target amplicon, until a match is found. If the distance between these flanking matches differs from the expected gap, the read will be screened for an indel and is not used to identify substitutions at the positions of interest. Once the analysis is complete, the somatic variant data are exported as two files. The first contains a report of the variants identified (Supplementary Figure S1), while the second contains the raw data showing the read depths for each nucleotide at each position of interest, as well as the number of reads that suggest the presence of an indel (Supplementary Figure S2). The AgileSMPoint program, a user guide and demonstration data are available from our website: http://dna.leeds.ac.uk/agile/AgileSMPoint/.

AgileSMAll: identification of somatic sequence variants at each position in an amplicon

As AgileSMAll screens the majority of positions within an amplicon for sequence variants, it is likely that it will detect sequence variants that are not physiologically important. Therefore, AgileSMAll not only reports the variant's nucleic acid substitution, but also identifies its location within a gene (ie, intronic, exonic or in a splice site). If the variant is in a coding sequence, the predicted amino-acid substitution is also reported. To identify these features, AgileSMAll requires a genome annotation file, which contains information on the location of all coding sequences in the genome. This file can be created by AgileSMAll, as described in the online user guide. The format of the amplicon target file is similar to the fasta file format, with the information for each amplicon present on two lines (Figure 2b). The first line contains the gene's name, a 'tab' character and the reference genome coordinate of the first (5′) nucleotide in the amplicon. The next line contains the sequence of the amplicon (excluding library adaptors) as it occurs on the (+) strand of the genome reference sequence used to create the genome annotation file, followed by the sequence of the amplicon's primer sequences (excluding library adaptors tags). Again, each sequence is separated by a tab character.

AgileSMAll identifies the origin of each read by comparing its 5′ sequence with the primer sequences described in the target file. As, however, a significant proportion of reads may derive from aberrantly amplified PCR products, AgileSMAll scans the first six bases following the primer sequence and ignores any sequence reads where this differs from the reference sequence. Next, AgileSMAll scans the remaining sequence to identify reads containing a possible indel. If a read may contain an indel, the start of the indel is found. Reads with indels at the same positions are combined to form a single read, which is then aligned to the reference sequence, allowing the indel's structure to be deduced. Only reads confirmed not to contain an indel are used to identify single base-pair substitutions.

Once AgileSMAll has read the entire contents of a data file, it identifies the presence of any variants. If the variant allele read depth is above the user-defined cutoff values, the variant is annotated and exported. AgileSMAll exports the data as three files: a report file, a raw data file and an indel alignment file. Examples of these files can be found in Supplementary Figures S3–5 in the Supplementary Data. The report file (Supplementary Figure S3) first describes the total number of reads mapped to each amplicon, the number that may (or may not) contain indels and the number of reads identified

as originating from an homologous sequence. Each description line of this type is followed by a list of the annotated variants ordered by position in the amplicon. Each line in this list describes a variant and shows the total number of reads with the variant base followed by the number of reads containing the reference base. If no variants were found, 'no mutations' are written below the amplicon description line. The indel alignment file (Supplementary Figure S4) contains a description of each indel followed by a comparison of the reference sequence and the typical sequence of the reads found to have an indel at the specified location. These data enable the user to identify indels that may be the result of sequencing error, particularly dealing with sequences containing mononucleotide repeat runs. The raw data file (Supplementary Figure S5) details the number of reads containing each of the four nucleotides at each position in the amplicon. As this file contains all the information necessary to detect substitution mutations in each amplicon, it is possible to use this file to reanalyze the data using different cutoff values, rather than re-reading the original, much larger sequence files. The AgileSMAll program, a user guide and demonstration data are available from our website: http://dna.leeds.ac.uk/agile/AgileSMAll/.

RESULTS

Twenty tumor samples were analyzed for *EGFR* mutations and a second set of 25 tumors for *BRAF* and *KRAS* mutations. These samples had previously been screened for somatic mutations at the known mutational hotspots in codon 600 of *BRAF*; codons 718, 744, 752, 767, 773, 789, 857 and 860 of *EGFR*; and codons 11, 12, 13 and 61 of *KRAS*. In this way, it was possible to compare the known genotypes with those identified using the AgileSMPoint and AgileSMAll programs.

Sensitivity of the Analysis

As AgileSMAll screens all the positions within a PCR product >6 bases from a primer, it detected a number of sequence variants not identified either by the earlier diagnostic screening or by AgileSMPoint (which only screened specified positions of known pathological importance). Most of these were present in 30% or more of the total reads, and were known SNPs. Interestingly, the remaining variants not identified by the previous analysis were present in only one sample of the *EGFR* cohort. This suggests that this sample may have had an intrinsically higher experimental error, possibly due to chemical modification of the DNA as a result of the formalin fixation process. Although AgileSMPoint and AgileSMAll identified all the variants detected by the prior diagnostic screening, AgileSMPoint also detected the presence of four extra variants occurring at known mutational hotspots and present in approximately 1% of reads. We did not attempt to distinguish whether these were artifacts of PCR or formalin fixation.

When the distribution of the proportion of base calls that differ from the reference sequence at each non-primer position in the *BRAF* and *KRAS* amplicons was examined (Supplementary Table S5 and Supplementary Figure S6), it could be seen that most positions were associated with non-reference sequence base calls. The distribution of these non-reference calls suggests that it will not be possible to discern if a low allele fraction variant ($<2\%$ of reads) identified by AgileSMPoint and AgileSMAll is a biologically genuine mutation, or the result of experimental artifact (created by formalin fixation, PCR error or sequencing error).

Optimum Read Depth

The *EGFR* and *BRAF* — *KRAS* data sets consisted of 125 (25×5) and 60 (20×3) amplicons, respectively. At this degree of multiplexing, very high read depths were obtained for all of the amplicons. As stated above, this does not necessarily increase sensitivity. Therefore, we performed a series of *in silico* experiments where the number of reads used in the analysis was reduced. These experiments suggested that it is possible to consistently identify variants at read depths of approximately 2000 reads. However, we found that this analysis was confounded by the difficulties of creating an equimolar pool of amplicons to be sequenced. As can be seen from Tables 1 and 2, the read depths vary several-fold between different amplicons. This suggests that difficulties in creating an equimolar pool of amplicons are a practical concern when choosing the number of samples to multiplex per lane. When pooling more than ~ 100 amplicons per lane, the time spent in equalizing the representation across samples becomes prohibitive, unless a robotic solution is used.

It can also be seen that the number of reads flagged as identifying a specific sequence variant differs between AgileSMPoint and AgileSMAll. This reflects their different approaches to identifying, which amplicon a read represents and whether or not a read originates from a pseudogene sequence. As there are no pseudogenes for *EGFR*, differences in read depths in this data set are a direct consequence of the method each program uses to identify the origin of a read. AgileSMAll detects a slightly higher number of reads per variant as it uses the 5′ part of a read to deduce its origin. This tends to have higher base-calling quality scores than the sequences used by AgileSMPoint. However, if primers of low synthesis quality and purity are used, the aberrant primer sequences in the amplicon hinder AgileSMAll's ability to identify its origin and can have a major effect on the read depth identified by AgileSMAll.

When screening the *BRAF* and *KRAS* data sets, which could also contain reads from pseudogene sequences, all the variants were found to have a very similar proportion of supporting reads. This suggests that both programs were equally effective at distinguishing reads originating from the pseudogenes. If the analysis was repeated using amplicon descriptions that lacked information on the divergent positions between the gene and pseudogene, the read depth at

Table 1 Variants identified in the BRAF (NM_004333.4) and KRAS (NM_004985.3) data by Agile2 and Agile1

	BRAF (1)[a]	Amplicon 2 (KRAS)		KRAS (3)[a]					
	c.1799T>A	c.182A>T	c.181C>T	c.38G>A	c.37G>T	c.35G>A	c.35G>T	c.34G>C	
Sample ID	V>E	Q>L	Q>K	G>D	G>C	G>D	G>V	G>C	WT
11-1						55% (43 727)			
						55% (52 904)			
11-52						54% (49 566)			
						54% (70 822)			
11-67		36% (49 738)							
		36% (51 557)							
11-108						37% (45 740)			
						36% (62 063)			
11-243				36% (58 753)				<5% (58 796)	
				36% (79 332)				2% (79 332)	
11-260				31% (44 363)					
				31% (68 112)					
11-295									WT
									WT
11-346	38% (64 521)								
	36% (110 083)								
11-457						33% (61 542)			
						33% (94 365)			
11-463	38% (115 707)								
	36% (189 143)								
12-4						14% (55 960)			
						14% (47 437)			
12-79							20% (57 883)		
							20% (86 088)		
12-102			<5% (109 994)			49% (121 694)			
			1% (108 286)			49% (172 725)			
12-166							46% (48 702)		
							46% (70 685)		
12-177						28% (89 568)			
						28% (127 589)			
12-219	21% (54 962)							<5% (61 791)	
	19% (98 420)							2% (87 704)	
12-238	51% (100 693)								
	49% (169 641)								
12-242	16% (119 479)								
	15% (208 984)								
12-268									WT
									WT
12-303					44% (41 213)				
					44% (44 671)				

For each variant, the proportion of reads indicating a variant is shown as a percentage of the total number of reads scored (shown in brackets) by Agile2 and Agile1, respectively. WT identifies samples found to contain no variants. Underlined cells identify variants reported in the diagnostic screening.
[a]The number in brackets indicates which amplicon contained each variant.

Table 2 Variants identified in the *EGFR* (NM_005228.3) data by AgileSMAll and AgileSMPoint

PCR product¥	EGFR (1)			EGFR (2)				EGFR (3)						EGFR (4)			EGFR (5)
Genomic position and substitution	c.2125G>A	c.2155G>T	c.2156G>C	c.2235A Del GGAATTAAGAGAAGC	c.2236Del GAATTAAGAGAAG>AGAC	c.2237Del AATTAAGAGAAGCAACAT C	c.2240Del TAAGAGAAGCAACATCTC	c.2300C>T	c.2301C>T	c.2302A>G	c.2303G>T	c.2309A>G	c.2309Ins CCAGCGTGG	c.2361G>A	c.2369C>T	c.2573T>G	c.2582T>A
Protein Change†	E>K	G>C	G>A	FS	FS	FS	FS	A>V	WT	S>G	S>I	D>G	FS	WT	T>M	L>R	L>Q
SNP	+	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-
1							39% (225479) 35% (164038)										
2			25% (115290) 24% (125583)					51% (48909) NA						100% (169900) NA			
3		42% (108417) 40% (119887)									51% (48421) 50% (155577)			58% (173369) NA			
4														100% (176346) NA			8% (191639) 7% (181642)
5														63% (165251) NA		15% (202121) 15% (194528)	
6											23% (38119) 23% (141227)					17% (223680) 17% (210569)	
7	42% (112223) NA													51% (140273) NA		42% (197089) 41% (187199)	
8														99% (158490) NA		41% (222125) 40% (207523)	
9														58% (197461) NA		37% (249521) 37% (223783)	
10					40% (911465) 69% (284574)									100% (147923) NA			
11						48% (342435) 46% (242406)								59% (138277) NA			

Table 2 (Continued)

12									5% (35706) NA	5% (35834) NA		5% (35787) NA				13% (161550) 12% (147543)	
13														37% (161584) NA	28% (162133) 30% (160492)	42% (209161) 41% (194524)	
14					72% (268736) 68% (161586)									75% (161719) NA			
15														51% (136609) NA		7% (303045) 6% (278106)	
16																31% (225451) 31% (210722)	
17														100% (202968) NA			
18														52% (114128) NA			
19														100% (157468) NA			
20													*5% (946223) 4% (192293)	100% (30347) NA			
21														100% (170651) NA			< 5% (54375) 1% (51171)
22			73% (88390) 72% (207565)											100% (130914) NA			
23														66% (135722) NA		43% (125192) 42% (117027)	
24														100% (125313) NA			< 5% (154104) 1% (145464)
25														53% (179163) NA		15% (119364) 15% (110896)	

For each variant, the proportion of reads indicating a variant is shown as a percentage of the total number of reads scored (shown in brackets) by AgileSMAll and AgileSMPoint, respectively. Shaded cells identify variants that had been reported in the previous diagnostic screening.
[a]The number in brackets indicates which amplicon contained each variant.
[b]FS and WT indicate frame shift and wild-type variants, respectively.
[c]The insertion variant g.55249011InsCCAGCGTGG was correctly reported only in reads originating from the forward strand, as the read depth for the variant on the reverse strand was below the 5% cutoff value used in this analysis.

each variant position noticeably increased, with a corresponding decrease in the proportion of reads supporting the variant. This suggested that both programs were discounting a large number of pseudogene-derived reads. Manual examination of the retained and discarded reads could not quantify the efficiency with which the reads were filtered, but the similarity of the variant read depth data reported by the programs when filtering out the pseudogene sequences suggested that both filtering mechanisms were robust.

Comparison of Somatic Variant Detection Using Aligned and Unaligned Sequence Data

When the sequence variant data sets produced by AgileSMall and AgileSMPoint are compared with the sequence variants identified using the BWA/VarScan pipeline (Supplementary Tables S6 and S7), it can be seen that BWA/VarScan detected all single base substitutions when the variant allele was present in $>5\%$ of the total number of reads. However, the BWA/VarScan pipeline did not identify any of the large indel variants present in the EGFR data set.

DISCUSSION

As tumor behavior is largely determined by patterns of somatically acquired mutation, there is considerable interest in cheap and efficient high-throughput methods that can quantifiably detect chosen examples of such variants. With the advent of massively parallel sequencing technologies, it has become comparatively trivial to generate the required amount of sequence data to detect somatic mutations in a quantifiable manner. However, owing to the volume of data, analysis has become a significant bottleneck in their detection. To simplify and speed up diagnostic analysis, we have adopted an amplicon-based NGS library production method and developed two novel programs to detect the presence of somatic mutations in tumor samples. When screening a comparatively small number of positions with AgileSMPoint, it is possible to identity variants present in as little as 1% of chromosomes in a sample. Owing to the greater number of positions screened by AgileSMAll resulting in more false positives, AgileSMAll can reliably identify variants present in 5% or more of the sequences in a sample. The effect of adjusting this cutoff value on the number of variants identified by AgileSMAll can be seen in Supplementary Table S8, for each data set. The most appropriate value for this cutoff should be determined by the user for each data set, as the number of false positives is affected by the method of FFPE fixing, amplicon sequence and PCR amplification protocol.

For the work described here, we chose to create the amplicon-based NGS libraries in a single PCR step. Although we have successfully used this method, as with all PCR-based approaches, it may require a degree of modification and optimization according to the nature of the target sequences. Typically, we have found the molar ratio of target-specific primers to the Illumina adapter oligonucleotides to be an important parameter when optimizing library production. In certain instances, it may be necessary to ligate the adaptor sequences to the amplicons in a second step.

Although the capabilities of AgileSMAll are broader than those of AgileSMPoint, both programs have specific features making them better suited to different scenarios. AgileSMPoint was designed to detect variants at known pathologically important positions in oncogenes, and so performs only a subset of the analysis that AgileSMAll is capable of. Although AgileSMAll is able to perform this simpler task, in doing so it would also identify variants of unknown significance ('VUS'), which at best would be ignored and at worst may cause a dilemma regarding the reporting of VUS. Consequently, by restricting the range of information reported, the AgileSMPoint program is more suited to diagnostic situations in which treatment decisions are directly based on genotype(s) at specified positions.

The greater range of variants detected by AgileSMAll makes it more suitable for the identification of deactivating mutations in tumor-suppressor genes, where important variants may occur in more loosely designated positions across the gene. In this situation, the detection of variants of unknown importance cannot be avoided and so protocols for assessing their importance must be developed. The greater capabilities of AgileSMAll may also make it more suitable for use in a research setting, where ethical and consent issues have normally been resolved at the start of the project and the detection of VUS may be a desired outcome.

When the sequence data were analyzed using the BWA/VarScan pipeline, all single-base variants with an alternative allele read depth greater than threshold of 5% of the total number of reads were identified. However, the large indel variants were missed. Although the BWA/VarScan pipeline may be tunable to improve the pickup rate of specific mutation types, we feel that this comparison with a commonly used set of command-line tools demonstrates the robustness of the AgileSMAll and AgileSMPoints analysis. Unlike AgileSMAll and AgileSMPoints, which ignored sequences derived from the PCR primers, the default BWA/VarScan pipeline screened the entire length of the amplicons, and so identified a large number of spurious variants derived from primer synthesis errors. As the sequencing target was generated by PCR amplification, the BWA/VarScan pipeline also identified variants present in sequences either originating from pseudogenes or aberrantly amplified by primer misannealing. Although such variants, if known, can be easily identified by their genomic position, the BWA/VarScan pipeline would need to be extended to filter out these variants, as they composed the majority of variants identified by this pipeline.

As discussed in the Results section, the available read depth for each PCR product does not limit the sensitivity of our methodology. Rather, the latter appears to be limited by artifactual sequence changes represented in the amplicons, as a result of either DNA damage caused by tissue fixation or

(for unfixed samples) PCR-induced errors. As formalin fixation is known to damage DNA, optimal sensitivity for detection of rare mutations is likely to be achieved by extracting DNA from fresh or frozen tissue.[20,21] However, given the current widespread diagnostic use of FFPE samples, we chose to use them to demonstrate the practical utility of our analysis methods with the sample types currently available.

Even when using DNA from unfixed fresh tissue, sensitivity may still be limited by the PCR-induced error rate and the sequencer's base-calling error rate. With NGS using Illumina technology, the manufacturer's criterion for a successful run requires a base-calling error rate of <1 in 1000 for only 80% of called bases; this implies that base-calling errors may be the most important factor limiting NGS-based somatic mutation detection. This is especially true for longer reads, when the quality scores for the later positions are worse than those for the earlier positions. The minimum achievable detection frequency for a minor allele has been reported in the literature to be between 3% and 0.1%[22–27] and as our data are derived from FFPE-treated DNA samples it is not surprising that our minimum allele frequency cutoff is toward the upper end of this range.

The maximum manageable level of sample multiplexing in a single experiment may be limited by the ability to create equimolar pools of PCR products, rather than by theoretical read depths. Creating equimolar pools of amplicons may seem a trivial task, but pipetting errors and difficulties in quantifying PCR product concentrations introduce a significant level of variation when pooling several hundred PCR products.

The detection of sequence variants may be confounded by the presence of SNPs in the amplicons or by the co-amplification of pseudogene sequences. It is particularly important to ensure that PCR primers do not anneal to sequences containing a SNP, as this could cause an amplification bias possibly excluding any variant on the same allele as the non-reference SNP allele. As AgileSMAll identifies the origin of a read by determining the primer used to amplify the amplicon, its analysis is not affected by the presence of SNPs in the amplified sequence. However, AgileSMPoint will only correctly analyze reads with SNPs close to the positions of interest if the program is instructed to do so in the target file (as described in the Materials and methods section). Both programs have mechanisms to distinguish reads originating from pseudogene sequences, with the divergent position highlighted in the program's target file. However, this depends on the divergent positions being present in each read. For paralogous sequences with no suitably placed divergent positions, it may not be possible to filter out pseudogene sequences. In this case, their presence should be taken into consideration when interpreting the results.

In conclusion, we have developed a robust practical methodology for the detection of somatic mutations down to proportions as low as 1% (AgileSMPoint) or 4% (AgileSMAll), using DNA extracted from FFPE tumor samples. The data analysis algorithms can detect and ignore sequences derived from co-amplified pseudogenes, while also correctly processing data containing known SNPs close to the positions of interest. The software applications (AgileSMPoint and AgileSMAll) are freely available from http://dna.leeds.ac.uk/agile/, where extensive user guides and demonstration data may also be found.

The simplicity of these methods for identifying rare sequence variants within a sample means that it is quite feasible to measure intratumor heterogeneity using multiple biopsies. As a result, as discussed above, variant detection is limited by factors other than read depth, the simplicity of our method opens up to analyze the important question of whether it is clinically more valuable to perform extensive analysis on a single sample per tumor or less exhaustive tests on a large number of samples from across the same tumor.

Supplementary Information accompanies the paper on the Laboratory Investigation website (http://www.laboratoryinvestigation.org)

ACKNOWLEDGMENTS

This work was supported by grants from Sir Jules Thorn Charitable Trust (Grant 09/JTA); Yorkshire Cancer Research (L354); EPSRC (Grant EP/K023845/1); the University of Leeds Mary and Alice Smith Memorial Scholarship and the MRC (MR/L01629X/1).

DISCLOSURE/CONFLICT OF INTEREST

The authors declare no conflict of interest.

1. McClellan M, Benner J, Schilsky R, *et al.* An accelerated pathway for targeted cancer therapies. Nat Rev Drug Discov 2011;10:79–80.
2. Brodeur GM. The involvement of oncogenes and suppressor genes in human neoplasia. Adv Pediatr 1987;34:1–44.
3. Carr IM, Camm N, Taylor GR, *et al.* GeneScreen: a program for high-throughput mutation detection in DNA sequence electropherograms. J Med Genet 2011;48:123–130.
4. Maher B. Exome sequencing takes centre stage in cancer profiling. Nature 2009;459:146–147.
5. Freed-Pastor WA, Prives C. Mutant p53: one name, many proteins. Genes Dev 2002;26:1268–1286.
6. Davies H, Bignell GR, Cox C, *et al.* Mutations of the *BRAF* gene in human cancer. Nature 2002;417:949–954.
7. Sasieni PD, Shelton J, Ormiston-Smith N, *et al.* What is the lifetime risk of developing cancer? The effect of adjusting for multiple primaries. Br J Cancer 2011;105:460–465.
8. Mamanova L, Coffey AJ, Scott CE, *et al.* Target enrichment strategies for next generation sequencing. Nat Methods 2010;7:111–118.
9. Srinivasan SM, Guda C. MetaID: a novel method for identification and quantification of metagenomic samples. BMC Genomics 2013; 14(Suppl 8):S4.
10. Leimena MM, Ramiro-Garcia J, Davids M, *et al.* A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics 2013;14:530.
11. Lange V, Böhme I, Hofmann J, *et al.* Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 2014; 15:63.
12. Cao H, Wang Y, Zhang W. A short-read multiplex sequencing method for reliable, cost-effective and high-throughput genotyping in large-scale studies. Hum Mutat 2013;34:1715–1720.
13. Chambers PA, Stead LF, Morgan JE, *et al.* Mutation detection by clonal sequencing of PCR amplicons and grouped read typing is applicable to clinical diagnostics. Hum Mutat 2013;34:248–254.

14. Koboldt DC, Zhang Q, Larson DE, *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–576.
15. Quach N, Goodman MF, Shibata D. *In vitro* mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. BMC Clin Pathol 2004;4:1.
16. Do H, Dobrovic A. Limited copy number-high resolution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies. Mol Cancer 2009;8:82.
17. Gallegos Ruiz MI, Floor K, Rijmen F, *et al*. EGFR and K-ras mutation-analysis in non-small cell lung cancer: comparison of paraffin embedded versus frozen specimens. Cell Oncol 2007;29:257–264.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760.
19. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009;25:2078–2079.
20. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am J Pathol 2002;161:1961–1971.
21. Williams C, Pontén F, Moberg C, *et al*. (1999) A high frequency of sequence alterations is due to formalin fixation of archival specimens. Am J Pathol 1999;155:1467–1471.
22. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 2010;26: i318–i324.
23. Fernald GH, Capriotti E, Daneshjou R, *et al*. Bioinformatics challenges for personalized medicine. Bioinformatics 2011;27:1741–1748.
24. Flaherty P, Natsoulis G, Muralidharan O, *et al*. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res 2012;40:e2.
25. Shendure J, Ji HP. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135–1145.
26. Vallania FL, Druley TE, Ramos E, *et al*. High-throughput discovery of rare insertions and deletions in large cohorts. Genome Res 2012; 20:1711–1718.
27. Wei Z, Wang W, Hu P, *et al*. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nuclei Acids Res 2011;39:e132.