SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
CANCER
DIAGNOSTIC MARKERS
BIOINFORMATICS

Received
2 October 2013

Accepted
6 November 2013

Published
3 December 2013

Computational optimisation of targeted DNA sequencing for cancer detection

Pierre Martinez[1], Nicholas McGranahan[1], Nicolai Juul Birkbak[2], Marco Gerlinger[1,3] & Charles Swanton[1,4]

[1]Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK, [2]Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark, [3]Barts Cancer Institute, Barts and The London School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK, [4]UCL Cancer Institute, Paul O'Gorman Building, Huntley St., London WC1E 6BT, UK.

Correspondence and requests for materials should be addressed to C.S. (Charles.swanton@cancer.org.uk)

Despite recent progress thanks to next-generation sequencing technologies, personalised cancer medicine is still hampered by intra-tumour heterogeneity and drug resistance. As most patients with advanced metastatic disease face poor survival, there is need to improve early diagnosis. Analysing circulating tumour DNA (ctDNA) might represent a non-invasive method to detect mutations in patients, facilitating early detection. In this article, we define reduced gene panels from publicly available datasets as a first step to assess and optimise the potential of targeted ctDNA scans for early tumour detection. Dividing 4,467 samples into one discovery and two independent validation cohorts, we show that up to 76% of 10 cancer types harbour at least one mutation in a panel of only 25 genes, with high sensitivity across most tumour types. Our analyses demonstrate that targeting "hotspot" regions would introduce biases towards in-frame mutations and would compromise the reproducibility of tumour detection.

Cancer research and biomedical sciences in general entered a new era with the *-omics* revolution. New technologies have permitted the study of cancer genomes together with their organisation and evolution at a depth never achieved before. The identification of driver genes by next-generation sequencing studies[1–4] the understanding of their role in tumorigenesis matched with efforts in drug discovery were anticipated to pave the way towards targeted therapies. However, personalised medicine still faces several challenges as there is yet no gold standard to robustly classify genomic aberrations as driver events[5,6], many cancer genes display evidence of context-dependent antagonistic function[7] and intra-tumour heterogeneity both fosters drug resistance and hampers biomarker development[8–13].

Accordingly, few biomarkers have been validated and are routinely used. There remains a need for non-invasive, more generalised methods applicable to early cancer detection as the majority of patients presenting with distant metastases at diagnosis still have poor overall survival. Improved methods of tumour detection, that would allow more patients to be treated before metastatic spread and with minimal disease burden, represent a vital area of research. Recent efforts in biomedical research have therefore focused on the analysis of circulating tumour cells and circulating tumour DNA (ctDNA)[14]. Extracting genetic tumour material from peripheral blood, or "liquid biopsy", is a non-invasive method of high potential for early diagnosis and therapeutic decision making[15–19], which has previously been used to monitor the evolution of resistance to *EGFR*-targeted therapy in colorectal cancer (CRC) through acquisition of *KRAS* mutations[20,21].

The potential of ctDNA for early detection is still to be determined, as it is not yet known how often mutations from the primary tumour can be reliably identified from the analysis of peripheral blood or how this would vary according to tumour stage. A recent study reported that mutated ctDNA could be detected in mice only a week after subcutaneous cancer cell injection and that it could be detected in over 50% of mice after 9 weeks using real-time PCR[22], suggesting ctDNA studies are approaching the sensitivity required to detect primary disease prior to imaging detection. The analysis of a cohort of 84 patients with paired plasma and formalin-fixed paraffin-embedded primary samples, spanning various cancer types, led to the detection of 62.5% of primary site mutations in ctDNA using the Sequenom MassArray System and OncoCarta panel[23]. Mouliere *et al.* investigated 38 CRCs using qPCR, detecting *KRAS* or *BRAF* mutations present in the primary tumour in all the paired plasma samples, including 4 stage II cases[24]. Interestingly, the frequencies reported in the stage II cases were not found to be lower than in higher stage cases. Furthermore, a multi-region analysis of 4 serous ovarian cancer cases revealed that only 18% of validated somatic mutations from the primary sites could be detected above background in the plasma via deep sequencing[25]. However, the plasma tended to be enriched for trunk mutations originating early in the developing pre-cancer clone and at least one trunk mutation could be reliably detected in each case.

Table 1 | Datasets

Tumour type (TCGA set)	Number of genes mutated in > 4% of discovery samples	Number of discovery samples	Number of TCGA validation samples	Number of non-TCGA validation samples	Percentage of early samples in TCGA	Average number of mutations in discovery samples	Average number of mutations in TCGA validation samples	Average number of mutations in non-TCGA validation samples
Breast (BRCA)	14	294	473	196	38.3	37.8	39.9	39.5
Colorectal (COAD,READ)	2603	272	217	70	55.6	383.8	192.7	407.7
Head & Neck (HNSC)	408	68	232	80	22.7	107.3	134.0	81.8
Kidney clear cell (KIRC)	32	185	122	17	60.3	63.7	63.6	23.3
Lung Adeno (LUAD)	1323	240	141	279	63.0	254.5	294.1	172.9
Lung Squamous (LUSC)	1110	133	43	NA	75.6	244.2	308.5	NA
Ovarian (OV)	106	27	431	8	5.9	52.7	45.1	30.5
Melanoma (SKCM)	3558	89	160	146	35.7	445.0	395.3	389.8
Thyroid (THCA)	11	74	229	NA	24.4	13.3	17.1	NA
Uterus (UCEC)	3211	180	61	NA	74.7	458.3	510.2	NA

Therefore, the combination of advanced sequencing technologies with tumour DNA analysis from peripheral blood may hold promise for early tumour detection. Since mutations present at primary tumour sites often represent only a small fraction of sequencing reads at a genomic position in ctDNA, a high sequencing depth is required for their reliable identification. Here, we investigate panels of limited genes across samples from various cancer types as a primary analysis to estimate the sensitivity achievable by cancer detection methods based on somatic mutations occurring in a targeted fraction of the genome. We find that up to 76% of all occurrences from 10 tumour types bear at least one mutation in a panel of only 25 selected genes, with high sensitivity in most specific tumour types. Our data further indicate that highly-targeted sequencing of "hotspot" regions would be more likely to miss out-of-frame mutations, which would hinder the reproducibility of the results in different cohorts.

Results

Discovery and validation cohorts. All mutation data were retrieved from the curated datasets published in[26]. To focus on mutations that may be detectable in early stage cancers, the TCGA samples were divided into a discovery cohort comprising stages I–II samples (early TCGA) and a validation cohort comprising stages III–IV samples (late TCGA). In order to avoid possible platform-specific biases that could arise from using only TCGA samples, mutations in each dataset from other published works were assembled, when available, in an independent non-TCGA validation cohort for 7 of the 10 tumour types. Ten different tumour types were represented by specific datasets and an additional "pan-cancer" set was created, regrouping all 10 types which account for 48% of all 2008 reported tumour occurrences[27]. Overall, the TCGA discovery cohort consisted of 1562 samples, the TCGA validation cohort of 2109 samples and the non-TCGA validation cohort of 796 samples (Table 1). The early stage TCGA samples of each set, used for discovery, represented 23% to 76% of all TCGA samples, with the exception of the ovarian set, of which only 6% were early stage samples.

Candidate genes. We computationally analysed the prevalence of mutations in the discovery and validation pan-cancer cohorts using panels of up to 25 genes. The sensitivity of targeted sequencing methods for early tumour detection was determined by the percentage of samples in the pan-cancer set bearing at least one mutation in each gene panel. Samples were weighted according to the occurrence of each tumour type and the number of samples in each set. We selected a maximum of 25 genes with elevated mutation rates as candidates by scoring genes according to the number of samples bearing a mutation, the number of mutations in each sample and the gene sizes (Table 2, see methods).

The tumour suppressor *TP53* was the best-ranked candidate, mutated in 30.7% of all discovery samples. Several other genes in the list are known tumour suppressors and oncogenes (*KRAS*, *PIK3CA*, *PTEN*, *VHL*, *FBXW7*, *SMAD4*, *APC*, *EGFR*) while the implication of others genes, such as *CDH10*, *DCAF4L2* or *PRDM9*, in tumorigenesis has yet to be determined. Weighting genes by size allowed the exclusion of large genes such as *TTN* and *MUC16* (respectively 107976 and 43524 bp, mutated in 44.0% and 24.3% of the discovery samples), where mutations are more likely to occur by chance alone and would be inadequate for targeted sequencing.

Mutational prevalence of candidate genes in discovery and validation cohorts. The candidate genes were analysed to find, for each possible number of genes in a combination (1 to 25), which combination corresponded to the highest achievable sensitivity (Figure 1A, Supplementary Table 2). Our analysis suggests that 76.1% of our discovery cohort could be identified by screening for mutations in only 25 genes, provided accurate detection methods and sufficient coverage. The sensitivity achievable in the TCGA validation cohort is

Table 2 | Mutation frequencies of the top 25 pan-cancer candidate genes in each tumour type (percentages of discovery samples)

Gene	Breast	Colorectal	Head & Neck	Kidney clear cell	Lung Adeno	Lung Squamous	Ovarian	Melanoma	Thyroid	Uterus	Pan-cancer	Weighted pan-cancer
TP53	25.2	44.5	47.1	1.6	42.9	68.4	55.6	13.5	0.0	16.1	30.7	38.0
KRAS	0.3	44.1	0.0	0.5	30.0	0.8	0.0	1.1	2.7	21.7	15.2	13.7
PIK3CA	31.3	45.6	11.8	2.7	5.0	9.8	0.0	2.2	0.0	51.1	22.3	20.9
PTEN	2.4	35.7	2.9	2.2	1.7	7.5	0.0	6.7	0.0	57.8	15.0	11.2
VHL	0.0	14.7	0.0	41.1	0.0	0.8	0.0	1.1	0.0	1.1	7.7	4.5
FBXW7	0.3	28.7	5.9	0.5	2.9	6.8	0.0	3.4	0.0	18.3	8.7	8.4
CDH1	5.4	15.4	2.9	0.5	0.8	0.8	0.0	2.2	0.0	3.9	4.7	4.9
CDH10	0.3	9.6	16.2	1.1	17.1	17.3	0.0	14.6	0.0	11.1	8.8	10.1
SMAD4	0.0	28.7	1.5	0.0	3.3	2.3	0.0	0.0	0.0	2.2	6.0	6.3
DCAF4L2	1.0	7.4	5.9	0.0	8.3	4.5	0.0	4.5	0.0	4.4	4.2	4.7
CTNNB1	0.3	19.5	1.5	0.5	3.8	1.5	0.0	3.4	0.0	30.6	8.0	6.0
OR2M3	0.0	5.9	0.0	0.0	5.8	3.8	0.0	10.1	2.7	6.7	3.7	2.9
FAM47A	0.7	4.8	2.9	0.5	9.2	7.5	0.0	14.6	1.4	6.1	4.8	4.4
FAM5C	0.7	8.8	5.9	3.2	11.7	12.0	3.7	21.3	0.0	5.0	7.0	6.8
PIK3R1	0.7	9.6	1.5	0.5	1.3	1.5	0.0	4.5	0.0	38.3	6.9	4.3
ZNF676	0.0	3.3	5.9	0.5	9.2	11.3	0.0	20.2	0.0	3.3	4.8	4.9
APC	0.3	77.6	1.5	0.5	3.3	3.0	0.0	6.7	0.0	13.9	16.5	16.0
KLHL4	0.3	8.1	7.4	0.5	8.3	5.3	0.0	9.0	1.4	7.2	5.0	5.4
WBSCR17	1.4	10.3	0.0	0.5	7.5	3.8	0.0	11.2	0.0	6.1	4.9	4.2
PRDM9	0.3	8.8	4.4	0.5	12.5	6.8	0.0	16.9	4.1	9.4	6.6	6.1
TPTE	0.3	7.0	4.4	0.0	13.3	10.5	0.0	23.6	0.0	8.9	6.8	6.2
PCDH11X	0.7	7.7	5.9	2.2	15.8	15.0	3.7	7.9	2.7	7.2	7.2	7.2
MAGEC1	0.3	7.7	5.9	1.1	9.2	9.8	0.0	15.7	1.4	9.4	6.1	5.8
EGFR	0.3	15.1	4.4	0.5	10.4	2.3	0.0	6.7	0.0	2.2	5.4	5.8
FAM47C	2.0	5.9	2.9	1.1	12.5	3.8	3.7	12.4	0.0	7.2	5.5	5.1

comparably high (76.7%), suggesting good reproducibility in higher stage tumours. The use of different experimental settings and sequencing techniques might explain the difference observed in the non-TCGA validation cohort, in which the sensitivity was 65.8%.

Figure 1B illustrates the relationship between achievable sensitivity and the quantity of DNA to be sequenced. The best combinations of candidate genes from the pan-cancer discovery cohort were defined for different thresholds of maximal nucleotide length, from 100 bp to the combined length of all 25 genes (60 kbp) using 100 bp increments (Supplementary Table 3). Mutations in a panel of five genes (TP53, KRAS, PIK3CA, PTEN, VHL), whose combined length is less than 7 kbp, are present in 61.2%, 63.0% and 49.3% of all cancers in the discovery, TCGA and non-TCGA validation sets respectively. Both graphs in Figure 1 also highlight that the achievable sensitivity follows a logarithmic-like curve, indicating that the addition of more candidate genes is unlikely to provide major improvements.

Specific cancer types. Figure 2 displays how each of the specific tumour types is represented by the candidate genes inferred from the pan-cancer set. When combining all 25 genes, 70 ± 24%, 71 ± 28% and 61 ± 25% of samples bore detectable mutations in the discovery, TCGA and non-TCGA validation cohorts respectively. Over 50% of samples harboured mutations in at least one of the pan-cancer candidate genes in 21 out of 27 cohorts (78%). Strikingly, recurrent mutations in colorectal and uterine adenocarcinomas and lung adeno and squamous cancers appear to be very well defined by the pan-cancer candidate genes, with a sensitivity above 80% achievable in the discovery cohort and above 90% in the TCGA validation cohort.

In contrast, thyroid cancer is the cancer type for which the somatic-mutation-based cancer detection would be the hardest, with only 14.9% of the discovery cohort presenting mutations in any of the pan-cancer candidate genes. The low number of mutations per sample in thyroid cancer (Table 1), along with the high frequency of *RET/PTC* rearrangements in this cancer[28], explain the poor predicted performance of somatic mutation scans in this tumour type. The low tumour detection in kidney clear cell (ccRCC) echoes the low proportion of *VHL* mutations in the dataset (114 out of 324 samples, including all cohorts), probably due to known technical issues in sequencing the 1st exon of the gene[1]. Supporting this contention, recent studies suggest *VHL* is expected to be mutated in over 80% of samples[1,29].

To determine the efficiency of having different tumour-specific gene panels rather than a global pan-cancer panel, the sensitivity that could be achieved using the pan-cancer candidate genes were compared to the one achievable using the best specific candidate genes for each tumour type (Figure 3, Supplementary Figure 2 and Supplementary Tables S4–S23). The strongest divergence was observed for ccRCC, indicating that the ccRCC driver genes are rarely contributing to the development of other cancer types. The median differences in sensitivity achievable using the pan-cancer and specific candidate genes were 9.8%, −0.2% and −2.9% in each cohort, thus only enhancing sensitivity in the discovery cohorts. This suggests that the pan-cancer candidate genes are only marginally sub-optimal compared with the best specific candidate genes in most tumour types. A gene panel regrouping all 190 specific candidate genes across all 10 tumour types would consist of 700,000 nucleotides. 90.5%, 93.2% and 86.6% of all cancers in the discovery, TCGA and non-TCGA validation cohorts respectively present at least one mutation in any of these 190 genes, suggesting this gene panel would only increase the sensitivity by 15 to 20% while sequencing 10 times as many nucleotides, thus increasing the sequencing cost 10-fold. These observations suggest that, given the recurrence of many cancer genes across different tumour types, targeting a panel of genes defined by the analysis of multiple cancer types might achieve a high overall tumour detection rate whilst still providing high sensitivity for most tumour types. Yet, the poor putative detection of thyroid cancer samples is a reminder that some cancers, in which distinct mechanisms are involved in tumorigenesis, would still require specific diagnostic methods.

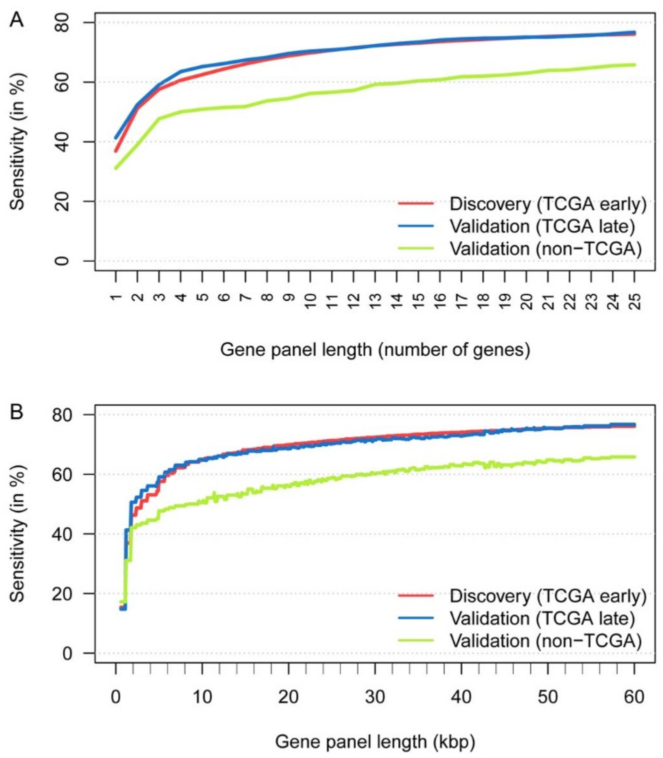

Figure 1 | Computationally estimated sensitivity of targeted sequencing screen for the detection of worldwide cancer occurrences. (A) Percentage of tumours in each cohort that have at least one mutation in a gene included in panels defined by combining the best 1 to 25 candidate genes. (B) Percentage of tumours in each cohort that have at least one mutation in a gene included in the best panel defined for a maximum length varying from 100 to 60,000 nucleotides.

Highly targeted scan: estimating the potential of hotspot regions. Recurrent mutations affecting a single nucleotide and resulting in activation or loss of function, known as "hotspots", often concentrate in certain regions of cancer genes, such as the DNA-binding domain of the *TP53* protein[30]. We thus investigated the presence of mutations in small genomic "hotspot regions", rather than whole genes, by grouping mutations close to each other into hotspot regions, using six different nucleotide distance thresholds (10, 20, 50, 100, 200 and 500 nucleotides, see methods). The selection of the best-ranked hotspot regions, up to a length equal to the combined nucleotide length of the pan-cancer candidate genes (60 kbp), revealed that this method could achieve sensitivities up to 94.6%, 81.9% and 67.4% in the discovery, TCGA and non-TCGA validation cohorts, respectively (Figure 4, Supplementary Tables S24–S29). Targeting hotspot regions might be capable of detecting a high number of cancers by sequencing less genetic material: 86.4%, 76.8% and 61.2% of cancers in the discovery and validation cohorts harbour mutations in 20 kbp of hotspot regions, compared to 69.9%, 68.7% and 56.0% using the best combination, whose length did not exceed 20 kbp, of the top 25 candidate genes (Figure 1B).

However, hotspot regions reveal a strong difference in sensitivity between the discovery and the validation cohorts: the percentage of cancers in the discovery cohort with detectable mutations is predicted to be on average 11.0% higher than in the TCGA validation cohort and 25.5% higher than in the non-TCGA validation cohort (Supplementary Tables S24–S29). Such differences are much higher than when using whole-gene panels. Furthermore, the analysis of mutation types in the entire sequence of mutated genes compared to the fraction covered by hotspot regions revealed that hotspot regions are enriched for in-frame mutations in the validation cohorts ($p < 0.001$ in 10 out of 12 cases, Fisher's exact test, Supplementary Figure 3). The contrary is observed in the discovery cohort, in which the mutations detectable by targeted sequencing of hotspot regions are depleted of in-frame mutations ($p < 0.001$ in all 6 cases). This indicates that highly-targeted methods, such as focusing on hotspot regions, are more likely to miss frameshift and truncating mutations occurring far from the point-mutation-rich active sites of many cancer genes, which would alter the reproducibility of tumour detection in different cohorts.

Single nucleotide variants. To investigate the sensitivity that could be achieved at single nucleotide resolution, we examined the recurrence of Single Nucleotide Variants (SNVs) in the pan-cancer dataset (Supplementary Table S30). Our analysis shows that screening for 1,000 unique single base pair substitutions could achieve sensitivities ranging between 41% and 73% of all occurrences from 10 tumour types (Figure 5). The 100 best ranked SNVs are estimated to be present in 44.6%, 35.9% and 28.2% of occurrences in the discovery and, TCGA and non-TCGA validation cohorts respectively. The high divergence in sensitivity between cohorts suggests that, as with hotspot regions, the sensitivity achievable through SNV screening would be highly dependent on the set used for discovery, which would hamper reproducibility. Yet, these results highlight the high potential of targeted SNV screen for early cancer detection. Especially high sensitivities are reported for colorectal and uterine cancers (Supplementary Figure 4), due to the prevalence of SNVs in the *KRAS* oncogene and overall higher mutational loads. Furthermore, the 1,000 SNV panel spans a high number of genes (765) and the sensitivity in the thyroid cancer samples is higher than with the 25 pan-cancer candidate gene panel.

Discussion

As the perspectives of personalised medicine are hindered by the heterogeneity found in individual tumours and the parallel evolution of subclones, the analysis of circulating tumour DNA represents an opportunity for major improvements in early diagnosis and tumour monitoring methods. Our analyses show that targeted screening methods have the potential to detect most cancers whilst limiting the amount of genomic DNA to be sequenced. We find that an estimated sensitivity of 65–77% could be achieved across 10 cancer types by sequencing only 25 genes, accounting for less than 60,000 nucleotide pairs. This represents approximately 0.002% and 0.2% of the human genome and exome respectively, indicating that targeted methods could provide a highly cost-efficient sequencing approach for cancer detection. We estimate that even greater sensitivity could be achieved through the sequencing of 190 genes, consisting of 700 kbp and representing 2.3% of the exome, which could increase sensitivity to 87–93%. In addition, a high sensitivity could be achieved in lung and colorectal cancers, suggesting great potential for early detection of these highly prevalent tumours. Coupled with the decreasing cost of next-generation sequencing techniques, the results presented here are encouraging in view of the increasing research-based use of peripheral blood circulating markers for cancer evolution analysis. However, sequencing errors can still be produced by current methods and thorough validation of mutation data is needed for more reliable sensitivity estimates.

Although several millions of mutations have been reported in thousands of sequenced tumours[26], it is becoming obvious that there are only few "mountains" in the mutational landscapes of tumours, with possibly as little as 140 genes significantly contributing to tumour development[31]. This provides a strong advantage for early pan-cancer detection by targeted sequencing. As branched evolution has been reported to occur in tumour development[8,11], panels of genomic regions for ctDNA screens for early detection should also be based on events likely to be involved in the early initiation of tumorigenesis (clonally dominant, trunk events) rather than somatic

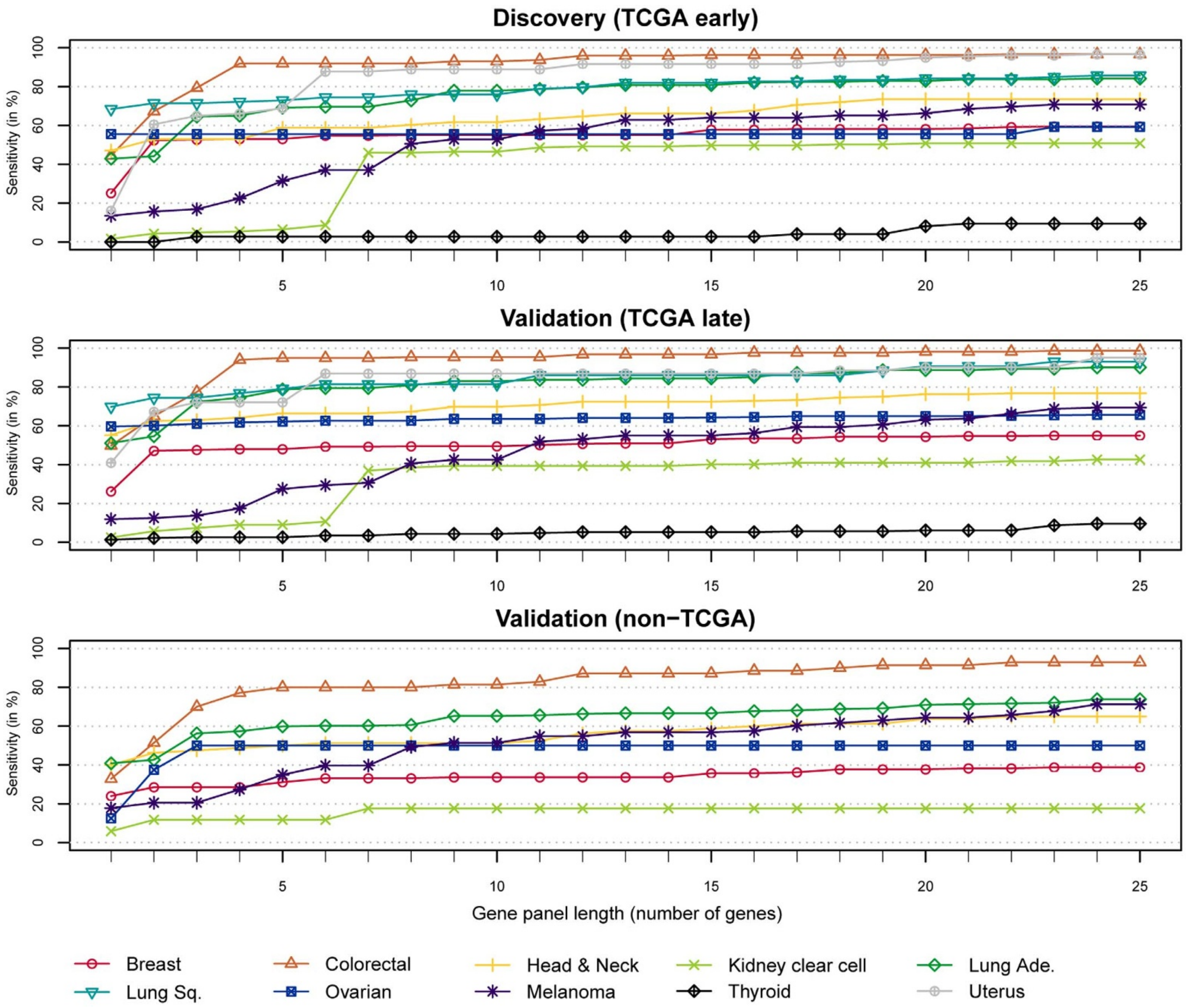

Figure 2 | Computationally estimated sensitivity of targeted sequencing screen for the detection of specific tumour types using pan-cancer candidate genes. Percentage of tumours for each tumour type in each cohort that have at least one mutation in a gene included in panels defined by combining the best 1 to 25 pan-cancer candidate genes.

events acquired later in tumour development (heterogeneous, branched events)[13]. Furthermore, previous studies were able to identify KRAS mutations in the blood of patients with colorectal cancers months before disease progression was detected by imaging[20,21], suggesting that peripheral blood-based techniques are sensitive enough to detect relatively small clonal populations.

There are still many hurdles prior to clinical application of targeted ctDNA analysis and the sensitivity of early diagnostic methods achievable in the clinic will greatly depend on the reliable detection of somatic mutations in ctDNA. The most important tasks will therefore be to assess how reliably tumour initiating somatic mutations which are present in primary tumour sites can be detected in ctDNA and to improve the current limitations to detection of ctDNA in patients. Deep sequencing appears to be a promising technique and can detect mutations above background[25] but further improvements are required to reduce error rates. Novel technologies based on redundant sequencing such as Tam-Seq[16], Safe-SeqS[32] or smMIP[33] can detect mutations with allele frequencies as low as 0.02% to 0.001% and already report up to 97% sensitivity for mutations with allele frequencies above 1%. Yet, the relationship between sensitivity, sequencing depth and tumour stage is unknown. Comparative studies with mutations detected in healthy controls will be essential in order to assess the specificity of plasma-based tools. Since driver genes are often mutated in many different tumour types is likely to complicate the identification of the original tumour site and additional methods, such as imaging, would be needed to bridge the gap between non-specific detection and adequate therapy.

Another unknown is the extent to which cancers and pre-cancers shed tumour DNA in the peripheral blood and if this is likely to differ according to tumour types. Additional studies are therefore needed to assess the potential of ctDNA-based analysis in each disease and determine the quantity of blood that would be necessary for reliable tests. In the case of cancers driven by somatic aberrations other than mutations, ctDNA scans could as well be tailored to detect gene fusions, promoter methylation or copy number changes, thus improving the potential for tumour detection. Our study further demonstrates that not all tumour types are equally well suited for somatic mutation-based scans, with colorectal cancer showing high estimated detection rates with good reproducibility whereas poor results are observed in thyroid cancers.

Provided efficient ctDNA scan solutions can be achieved, the next step would be the development of algorithms to establish the most

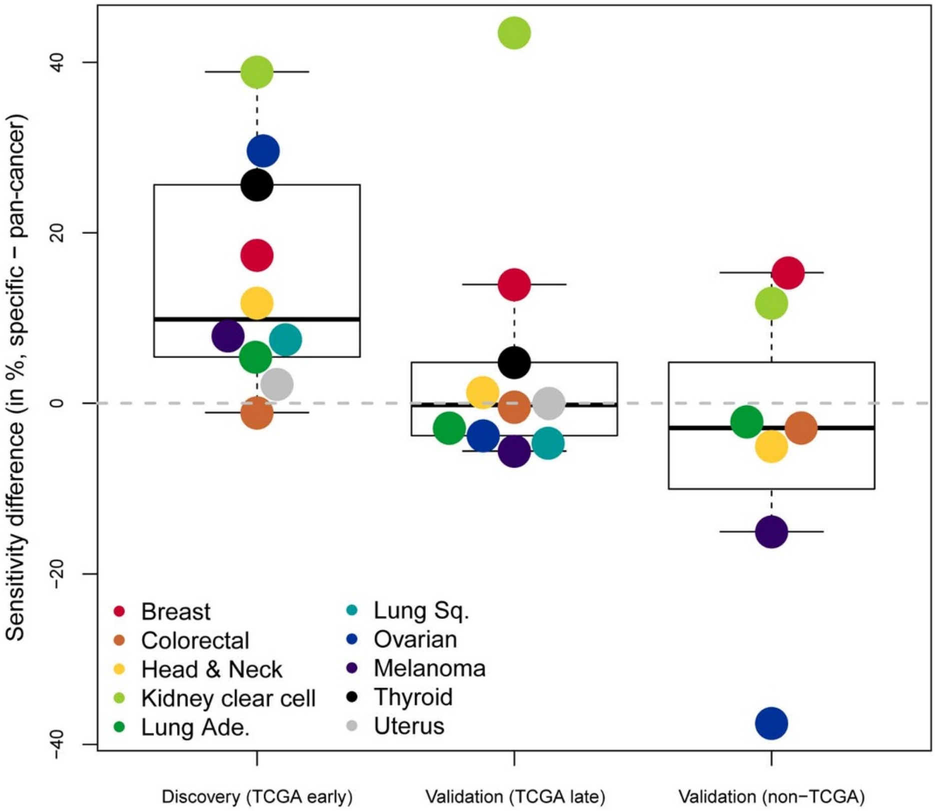

Figure 3 | Differences between pan-cancer and specific candidate gene panels in computationally estimated sensitivity of targeted sequencing screen for tumour detection. Coloured dots illustrate the difference between the sensitivity estimated using pan-cancer candidate genes and the sensitivity estimated using specific genes for each tumour type. Boxplots illustrate the distributions of differences in each cohort, thick black lines highlight the median, boxes delimit quartiles and whiskers indicate 95% confidence intervals.

favourable gene/loci panels. Here, we used simple algorithms coupled with empirical thresholds to extract a limited number of candidate genes, suggesting that high sensitivity could be achieved in most tumour types by sequencing 40 to 500 times less genetic material, which might reduce sequencing cost by a similar factor. Improved computational methods could increase the search space and find more efficient combinations of genes or genomic loci. Our findings also suggest that targeting highly mutated "hotspot" regions would limit the number of base pairs to be sequenced but would be more likely to miss loss of function mutations occurring in tumour suppressor genes. Approaches focusing on SNVs would potentially suffer from the same limitations. Our analysis however demonstrates that 28% to 44% of occurrences of the 10 studied tumour types present at least one SNV from a panel of 100 and that 41% to 73% sensitivity could be achieved using a panel of 1,000 SNVs. More generally, there is a need for new bioinformatic tools to be developed that can facilitate clinical applicability, and these should be developed with a focus on finding optimal solutions to the sequencing cost to tumour detection ratio problem.

The results we present here suggest that targeting a small number of genomic loci could allow the early detection of a high number of cancers across multiple tumour types. This stresses the importance of leading tumour-type-specific studies, using paired primary and plasma samples, aimed at defining the sequencing depth required to reliably identify the mutations present in ancestral clones at different tumour stages. We suggest that the development of innovative bioinformatic methods could help design cost-efficient gene panels that would allow the detection of a large number of tumours whilst optimising both the length and number of DNA sequences to be screened.

Methods

Datasets. Mutation data were obtained from the curated datasets provided in the supplementary data of Alexandrov *et al.*[26], regrouping large-scale sequencing studies from multiple sources for different tumour types (ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/, "raw" files). To comply with the TCGA publication guidelines, only sets corresponding to stage I–IV solid tumour types with no restrictions of use after August 2013 from TCGA were included in the analysis (Breast: BRCA; Colorectal: COAD,READ; Kidney clear cell: KIRC; Head & Neck: HNSC; Lung adenocarcinoma: LUAD; Lung squamous: LUSC; Ovarian: OV; Melanoma: SKCM; Thyroid: THCA; Uterus: UCEC). Clinical data were downloaded from the TCGA website. Mutations were annotated using ANNOVAR[34] and hg19 genome annotations; those corresponding to known entries in dbSNP 132[35] were removed in order to eliminate likely germ-line SNPs, and only non-synonymous mutations were considered. Each set was split three ways: a discovery cohort consisting of early stage (I–II) TCGA samples, a validation cohort consisting of late stage (III–IV) TCGA samples and a second independent validation cohort consisting of non-TCGA samples of any stage (when available). The decision to use early stage samples as the discovery cohort was driven by the necessity to identify mutations that can be used to detect tumours at an early stage, assuming that mutations present at an early stage will also be present at a later stage in the absence of treatment. An all-inclusive set, labelled "pan-cancer" set, was created by regrouping all samples of all 10 tumour types and was similarly split.

Each cohort from each set was represented as a 2D mutation matrix *M* of genes per sample, consisting of 0 (no mutation of a given gene in a given sample) and 1 (mutation) values. In the case of the pan-cancer set, samples were further weighted by multiplying by the occurrence of each cancer type, as given by the GLOBOCAN 2008 study[27] (Supplementary Table 1), and dividing the number of samples of each type in the set to account for overall worldwide cancer occurrences.

Candidate genes selection. In each set, only genes with non-synonymous mutations in at least 4% of samples were analysed (4% of all weighted occurrences for the pan-cancer set, see "Datasets" section above). Each matrix column, corresponding to a sample, was divided by the total number of mutations in this sample to give less importance to genes recurrently mutated in samples with high mutational burdens. A score *S* was attributed to each gene *g* such that *Sg = sum(M[g,])/Lg*, where *M[g,]* are the values for gene *g* in each column (sample) of *M* and *Ls* the length of *g*. Gene lengths were defined as the nucleotide length of the longest protein coding sequence retrieved

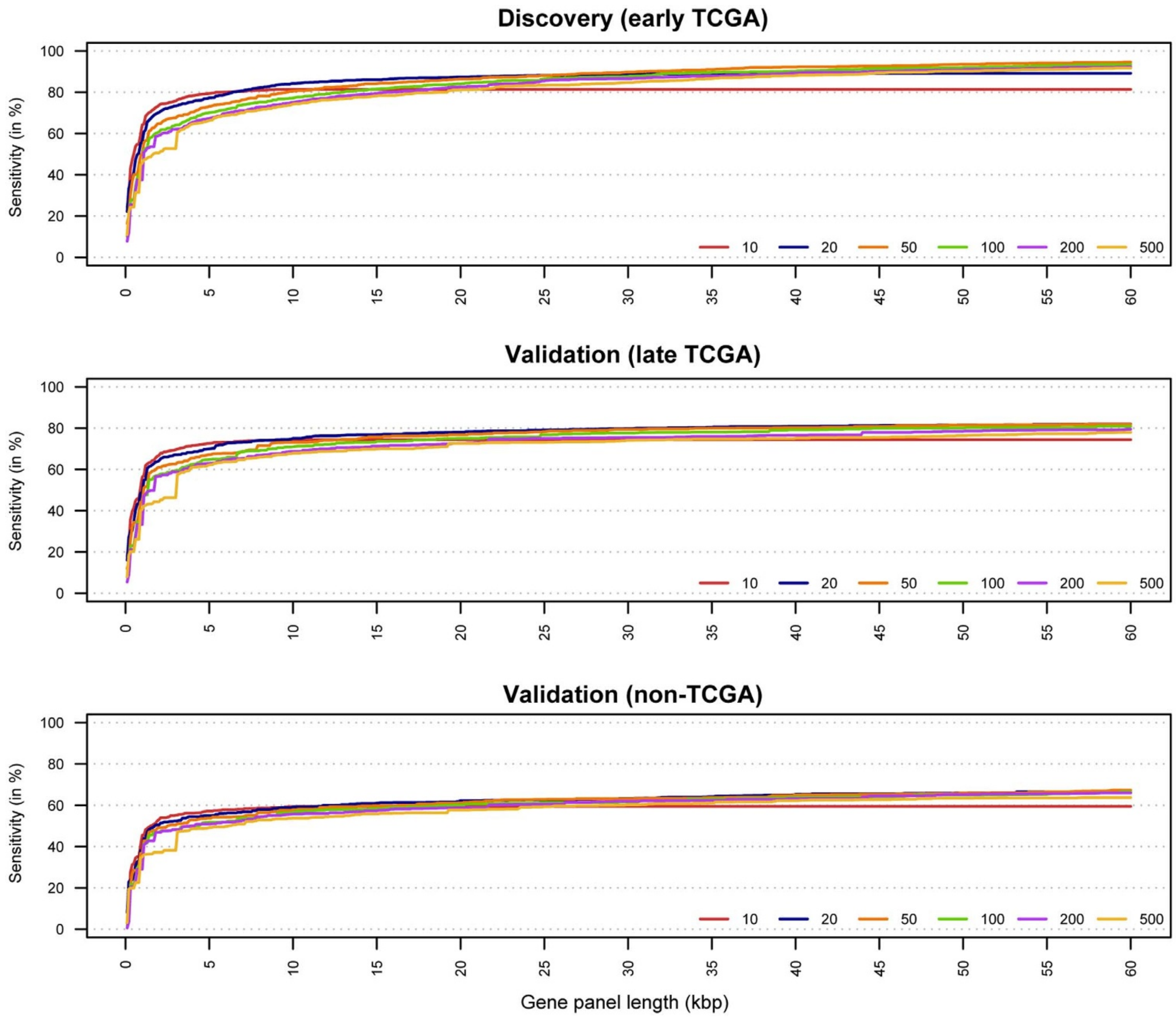

Figure 4 | Sensitivity and number of sequences to be sequenced in a highly targeted approach. Percentage of tumours in each cohort that have at least one mutation in any of the best-ranked hotspot regions for a maximum length of 100 up to 60,000 nucleotides. Coloured lines correspond to different nucleotide distances used to define hotspot region lists.

for each gene and the number of exons in a gene was determined as the highest number of exons in a single transcript of the maximum length. Both gene length and number of exons were retrieved from Ensembl (GRCh37.p11). For each dataset, a list of at most 25 candidate genes was defined using the genes with the highest *Sg* score. Only 14 and 11 genes were selected in the BRCA and THCA discovery sets respectively, given the limited number of recurrently mutated genes in these cancer types. All possible combinations of candidate genes were analysed to find the highest sensitivity, as given by the highest proportion of samples bearing a mutation in at least

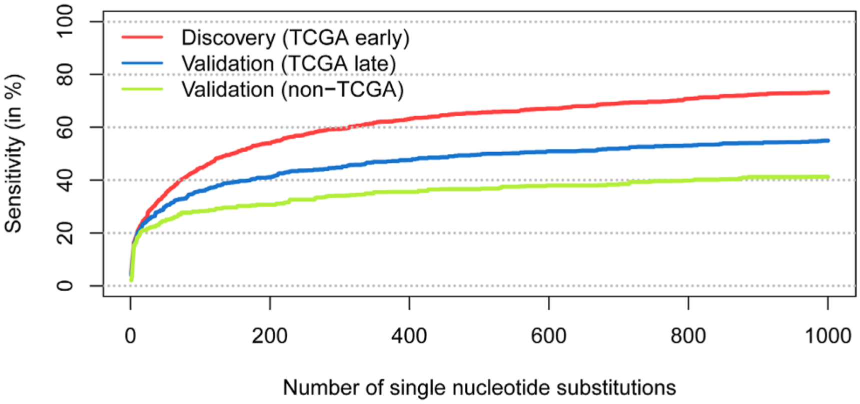

Figure 5 | Computationally estimated sensitivity of targeted SNV sequencing screen for the detection of worldwide cancer occurrences. Percentage of tumours in each cohort that have at least one mutation in panels defined by iteratively combining the best 1,000 single nucleotide variants.

one of the considered genes. In the case of the pan-cancer set, samples were weighted to reflect the worldwide occurrence of each tumour type.

Hotspot regions. Hotspot regions were defined by iteratively grouping mutations from the pan-cancer set in the discovery cohort that were at most *d* nucleotides away from one another. The possible threshold values of the distance *d* were 10, 20, 50, 100, 200 and 500 nucleotides (Supplementary Figure 1). Hotspot regions whose mutations could be detected in less than 5 samples were removed and a 2D mutation matrix *Mn* of hotspot region per sample was created for each value of *d*. Similarly to the previously described *Sg* score, a score *Sc* was computed for each hotspot region *h* such that *Sh = sum(Mn[h,])/Lh* using the region length *Lh* instead of the gene length and the hotspot regions were then sorted by *Sh* score. As with the pan-cancer candidate genes, samples were weighted to account for overall worldwide cancer occurrences. Different hotspot regions can represent different sections of a single gene and investigating all combinatorial possibilities would be very demanding computationally. Instead, for each threshold *t* corresponding to a length between 100 bp and 60 kbp (100 bp increments), a list of hotspot regions was created by iteratively adding regions until the length of the list reached *t*. All mutations in all genes at least partially mapping to the top 1,000 hotspot regions of each list constituted the background for in-frame mutation enrichment analyses; only the mutations exclusively comprised in the top 1,000 regions were considered as detectable, all others were considered as not detectable. A two-tailed Fisher's exact test was used to assess the enrichment of in-frame mutations in the mutations detectable by hotspot region sequencing.

Single nucleotide variants. Single nucleotide variants (SNVs) were defined as the unique 1 base pair substitutions occurring in the pan-cancer discovery cohort. This means that mutations of the same nucleotide at a certain genomic location (reference) to more than one different nucleotide (variant) will be considered as different SNVs. Those occurring more than once were summarized in a 2D matrix *Ms* of SNVs per sample. Similarly to whole genes and hotspot regions, samples were weighted by type to represent worldwide occurrences and by number of mutations. Each SNV *s* was given a score *Ss* given by *Ss = sum(Ms[s,])*. SNVs were sorted per *Ss* score and the best 1,000 were analyzed by iterative addition into a list, as for hotspot regions.

1. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542, doi:10.1038/nature09639 (2011).
2. The Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615, doi:10.1038/nature10166 (2011).
3. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764, doi:10.1038/ng.2291 (2012).
4. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472, doi:10.1038/nature09837 (2011).
5. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* doi:10.1038/nature12213 (2013).
6. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169–e169, doi:10.1093/nar/gks743 (2012).
7. Stepanenko, A. A., Vassetzky, Y. S. & Kavsan, V. M. Antagonistic functional duality of cancer genes. *Gene* doi:10.1016/j.gene.2013.07.047 (2013).
8. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892, doi:10.1056/NEJMoa1113205 (2012).
9. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313, doi:10.1038/nature10762 (2012).
10. Landau, Dan A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152**, 714–726, doi:10.1016/j.cell.2013.01.019 (2013).
11. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007, doi:10.1016/j.cell.2012.04.023 (2012).
12. Gerlinger, M. & Swanton, C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br. J. Cancer* **103**, 1139–1143, doi:10.1038/sj.bjc.6605912 (2010).
13. Martinez, P. *et al.* Parallel evolution of tumour subclones mimics diversity between tumours. *J. Pathol.* **230**, 356–364, doi:10.1002/path.4214 (2013).
14. Dennis Lo, Y. & Chiu, R. W. Plasma nucleic acid analysis by massively parallel sequencing: pathological insights and diagnostic implications. *J. Pathol.* **225**, 318–323, doi:10.1002/path.2960 (2011).
15. Dawson, S.-J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209, doi:10.1056/NEJMoa1213261 (2013).
16. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra168–136ra168, doi:10.1126/scitranslmed.3003726 (2012).
17. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* doi:10.1038/nature12065 (2013).
18. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* doi:10.1038/nrclinonc.2013.110 (2013).
19. Sozzi, G. *et al.* Quantification of free circulating DNA as a diagnostic marker in lung cancer. *J. Clin. Oncol.* **21**, 3902–3908, doi:10.1200/jco.2003.02.006 (2003).
20. Diaz, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540, doi:10.1038/nature11219 (2012).
21. Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532–536, doi:10.1038/nature11156 (2012).
22. García-Olmo, D. C., Picazo, M. G., Toboso, I., Asensio, A. I. & García-Olmo, D. Quantitation of cell-free DNA and RNA in plasma during tumor progression in rats. *Mol. Cancer* **12**, 8–8, doi:10.1186/1476-4598-12-8 (2013).
23. Perkins, G. *et al.* Multi-purpose utility of circulating plasma DNA testing in patients with advanced cancers. *PLoS One* **7**, e47020–e47020, doi:10.1371/journal.pone.0047020 (2012).
24. Mouliere, F. *et al.* Circulating Cell-Free DNA from Colorectal Cancer Patients May Reveal High KRAS or BRAF Mutation Load. *Transl. Oncol.* **6**, 319–328 (2013).
25. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* doi:10.1002/path.4230 (2013).
26. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259, doi:10.1016/j.celrep.2012.12.008 (2013).
27. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **127**, 2893–2917, doi:10.1002/ijc.25516 (2010).
28. Jhiang, S. M. The RET proto-oncogene in human cancers. *Oncogene* **19**, 5590–5597, doi:10.1038/sj.onc.1203857 (2000).
29. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867, doi:10.1038/ng.2699 (2013).
30. Bullock, A. N. & Fersht, A. R. Rescuing the function of mutant p53. *Nat. Rev. Cancer* **1**, 68–76, doi:10.1038/35094077 (2001).
31. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558, doi:10.1126/science.1235122 (2013).
32. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9530–9535, doi:10.1073/pnas.1105422108 (2011).
33. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854, doi:10.1101/gr.147686.112 (2013).
34. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164, doi:10.1093/nar/gkq603 (2010).
35. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

Acknowledgments

This work was funded by the European Union Framework 7 program PREDICT Consortium and Cancer Research UK. CS and MG are funded by the UK Medical Research Council. CS is funded by the Prostate Cancer Foundation and the Breast Cancer Research Foundation. The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Author contributions

P.M. and N.M. designed the experiments. P.M. performed all analyses and prepared all figures. P.M., N.M., N.J.B., M.G. and C.S. interpreted the data. C.S. supervised the work. P.M. and C.S. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Martinez, P., McGranahan, N., Birkbak, N.J., Gerlinger, M. & Swanton, C. Computational optimisation of targeted DNA sequencing for cancer detection. *Sci. Rep.* **3**, 3309; DOI:10.1038/srep03309 (2013).