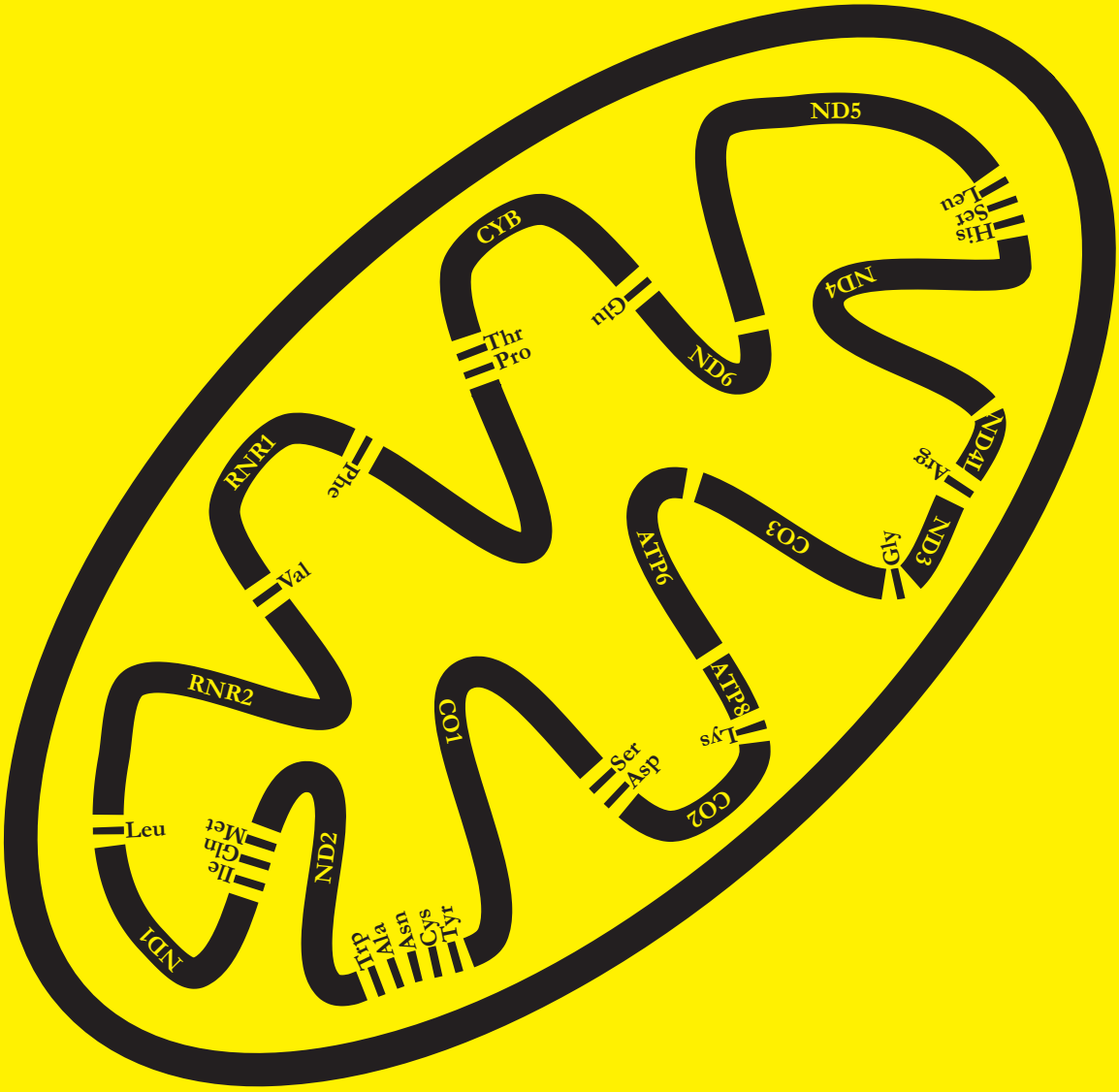
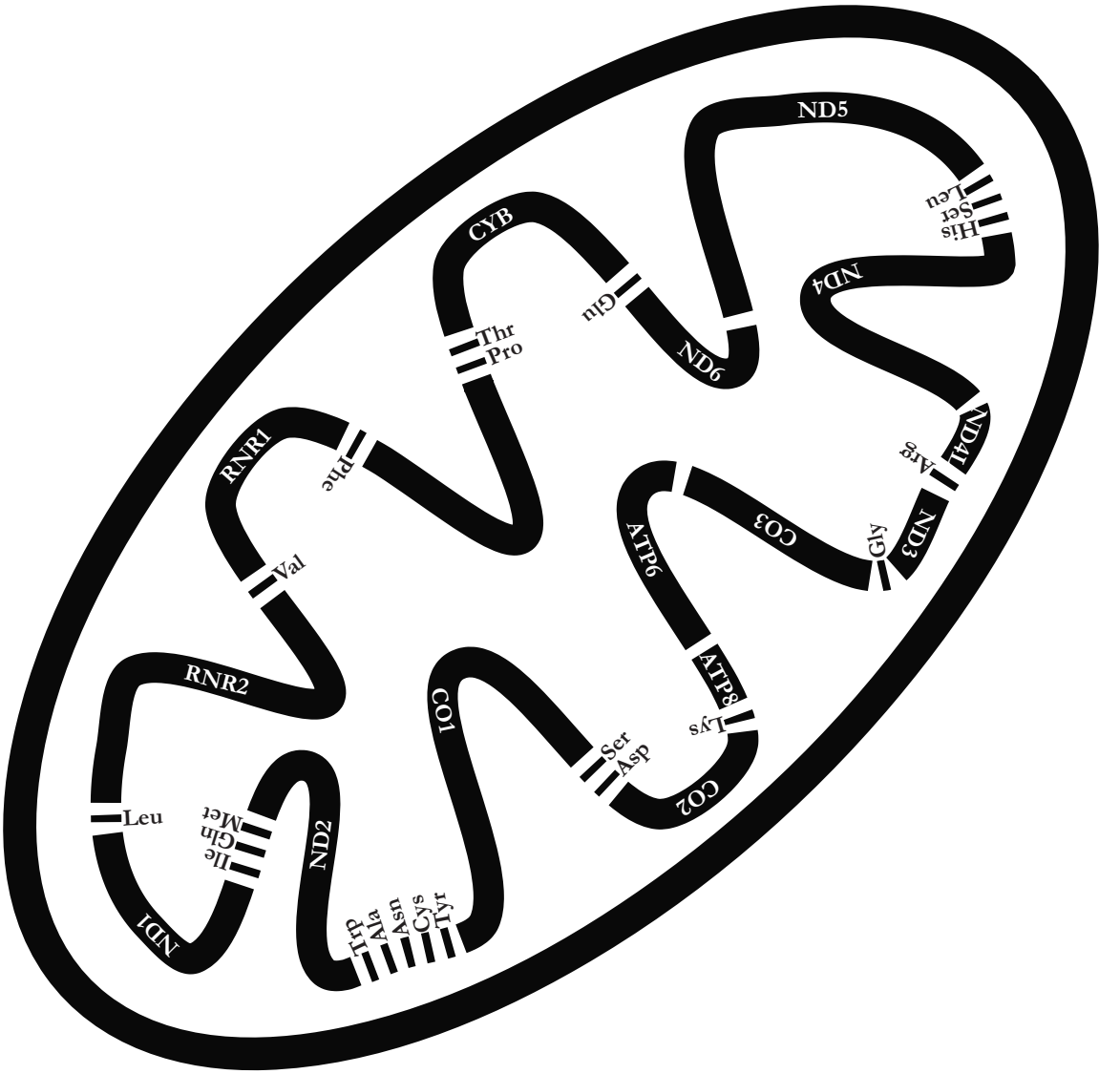
Mitochondrial DNA as a Breast Cancer Biomarker



Marjolein J.A. Weerts

Mitochondrial DNA as a Breast Cancer Biomarker



Marjolein J.A. Weerts

The studies described in this thesis were performed within the framework of the Erasmus MC Molecular Medicine (MolMed) Graduate School at the department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands.

The work described in this thesis has been carried out at the Erasmus MC under research agreement as part of a Philips Research program.

Financial support for printing this thesis was generously provided by: Department of Medical Oncology of the Erasmus MC Cancer Institute, Erasmus University Rotterdam, and the Molecular Pathway Diagnostics (Philips).

Cover	Marjolein Weerts
Layout	Renate Siebes | Proefschrift.nu
Printed by	Proefschriftmaken.nl
ISBN	978-94-6380-180-5

Mitochondrial DNA as a Breast Cancer Biomarker

Mitochondriaal DNA als bio-indicator bij borstkanker

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
dinsdag 19 februari 2019 om 15.30 uur

door
Marjolein Johanna Antonia Weerts
geboren te Venray

Erasmus University Rotterdam

Promotiecommissie:

Promotoren:	Prof.dr. S. Sleijfer
	Prof.dr. J.W.M. Martens
Overige leden:	Prof.dr. G.W. Jenster
	Prof.dr. R.M.W. Hofstra
	Prof.dr. H.J.M. Smeets

Over the long term, symbiosis is more useful than parasitism. More fun, too. Ask any mitochondria.

Larry Wall (?)

CONTENTS

CHAPTER 2


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Ser
Asp
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Ser
Leu

Mitochondrial DNA content in breast cancer: Impact on *in vitro* and *in vivo* phenotype and patient prognosis

Marjolein J.A. Weerts | Anieta M. Sieuwerts | Marcel Smid |
Maxime P. Look | John A. Foekens | Stefan Sleijfer | John W.M. Martens

Oncotarget 2016; 7:29166-29176

Abstract

Reduced mitochondrial DNA (mtDNA) content in breast cancer cell lines has been associated with transition towards a mesenchymal phenotype, but its clinical consequences concerning breast cancer dissemination remain unidentified. Here, we aimed to clarify the link between mtDNA content and a mesenchymal phenotype and its relation to prognosis of breast cancer patients. We analysed mtDNA content in 42 breast cancer cell lines and 207 primary breast tumour specimens using a combination of quantitative PCR and array-based copy number analysis. By associating mtDNA content with expression levels of genes involved in epithelial-to-mesenchymal transition (EMT) and with the intrinsic breast cancer subtypes, we could not identify a relation between low mtDNA content and mesenchymal properties in the breast cancer cell lines or in the primary breast tumours. In addition, we explored the relation between mtDNA content and prognosis in our cohort of primary breast tumour specimens that originated from patients with lymph node-negative disease who did not receive any (neo)adjuvant systemic therapy. When patients were divided based on the tumour quartile levels of mtDNA content, those in the lowest quarter (≤350 mtDNA molecules per cell) showed a poorer 10-year distant metastasis-free survival than patients with >350 mtDNA molecules per cell (HR 0.50 [95% CI 0.29-0.87], P = 0.015). The poor prognosis was independent of established clinicopathological markers (HR 0.54 [95% CI 0.30-0.97], P = 0.038). We conclude that, despite a lack of evidence between mtDNA content and EMT, low mtDNA content might provide meaningful prognostic value for distant metastasis in breast cancer.

Introduction

Mitochondria play a role in many cellular processes including oxidative phosphorylation, redox homeostasis, controlling calcium levels for regulation of signal transduction pathways, and the intrinsic apoptotic pathway [1]. The mitochondria contain their own genome – mtDNA – encoding their own translational machinery and 13 crucial proteins for the oxidative phosphorylation system. Related to energy needs, numerous mtDNA molecules may exist in a single cell. This number is not only dependent on the amount of mitochondria per cell but also on the number of mtDNA molecules per mitochondrion. Broad ranges in mtDNA content have been reported, from a few molecules in embryonic and pluripotent stem cells [2, 3] up to several thousands in subcutaneous adipocytes [4] or cardiac myocytes [5]. The cell-specific mtDNA content is assumed to be fairly stable under physiological conditions but can be altered by stress such as exogenous toxins [6], viral infection [7] and by genetic mutations [8]. The effects of changes in mtDNA content are illustrated in several mtDNA depletion syndromes [9], which are all characterized by impaired energy production.

Several studies examined mtDNA content in the context of cancer but so far no clear picture has emerged. In preclinical models, depletion of mtDNA yielded both increased and decreased *in vitro* tumorigenic phenotypes [10-17]. The *in vivo* findings using mouse xenografts are indecisive as well, as both gain and loss of tumorigenic potential upon mtDNA depletion has been reported [16-21]. Additionally, contradictory findings have been described for mtDNA content in human tumour specimens compared to their healthy counterparts in multiple cancer types (as reviewed in [22, 23]).

With regard to breast cancer, the impact of the mtDNA content on phenotype, prognosis and drug response has been investigated in several studies. Lower mtDNA content is observed in approximately 70% of breast cancer specimens when compared to their surrounding normal epithelium [24-31]. There are indications that low mtDNA content in breast cancer may yield a more aggressive phenotype and altered therapy responses. First, depletion of mtDNA in *in vitro* models affects the mRNA and protein expression levels of several genes involved in epithelial-to-mesenchymal transition (EMT) [12, 14]. The transition towards the mesenchymal phenotype has been implied as an essential mechanism contributing to cancer dissemination [32]. Consequently, low mtDNA content as a marker for the mesenchymal phenotype potentially identifies tumour aggressiveness. Second, a link between reduced mtDNA content and resistance to anti-estrogen regimens has been established in *in vitro* models [33]. Nevertheless, no association between estrogen receptor status and mtDNA content was observed in breast tumours [24-29]. Also, reduced mtDNA content was linked to a shift in drug response for breast cancer cell lines [17, 24, 34]. An *in vitro* reduction in mtDNA content revealed

increased sensitivity to cisplatin [17] and doxorubicin [24], but also decreased sensitivity to vincristine, paclitaxel and – in contrast to a previous study – doxorubicin [34]. In a small patient cohort, low mtDNA content was associated with longer disease-free survival in patients receiving adjuvant chemotherapy, whereas this was not the case for patients not receiving adjuvant treatment [24]. Few additional studies reported on breast cancer patient disease free- or overall survival in relation to tumorous mtDNA content [25-27]. However, these studies had either relatively small sample sizes or no information about treatments administered, the mtDNA content determination methods varied, and results were inconclusive.

Here, we further explore the putative link between mtDNA content and prognostic features in breast cancer. In a broad panel of human breast cancer cell lines the link between mtDNA content and a mesenchymal phenotype was studied by correlating it with expression levels of EMT-related genes and with the intrinsic subtypes of breast cancer [35, 36]. In a well-defined patient cohort of primary breast tumour specimens [37], tumour mtDNA content was examined in relation to expression levels of EMT-related genes, to the intrinsic subtypes, as well as to established clinicopathological variables. Primarily, in our cohort of primary breast cancer patients with lymph node-negative disease who did not receive any (neo)adjuvant systemic therapy, we examined the prognostic value of mtDNA content using distant metastasis-free survival as the main endpoint.

Results

mtDNA content in breast cancer cell lines and primary tumour specimens

In total, we analysed DNA extracts from 42 breast cancer cell lines and 207 primary tumour specimens. Multiplex real time quantitative PCR (qPCR) targeting a nuclear-encoded and a mitochondrial-encoded gene combined with array-based copy number changes of the nuclear-encoded gene to correct for sample specific somatic variation at the reference locus was used to obtain the mtDNA content in the DNA extracts of these samples. Inter-assay variability of the multiplex qPCR assay was monitored using the calibration curves taken along in each run ($n = 7$). Amplification in the calibration curve samples was linear between 0.16 and 16 ng DNA per reaction with mean efficiencies and standard error of 97.6 ± 4.4% for nuclear encoded *HMBS* and 91.5 ± 5.2% for mitochondrial encoded *MT-TL1*. Copy number variation of the nuclear encoded *HMBS* gene was observed in 39% of the breast cancer cell lines including 1 with homozygous loss, 12 with heterozygous loss and 3 with gain, and in 14% of the primary tumour specimens including 20 with heterozygous loss and 10 with gain. Because of a homozygous *HMBS* loss in SUM1315MO2, this cell line was excluded from further analysis. Because

of absence of *HMBS* qPCR signal amplification in three primary tumour specimens, these samples were excluded from further analysis as well. The median mtDNA content and interquartile ranges (IQR) in the 41 breast cancer cell lines and in the 204 primary breast tumour specimens were respectively 489 (IQR 360) and 462 (IQR 294) mtDNA molecules per cell.

Table 1 mtDNA content in the intrinsic breast cancer subtypes.

	Subtype	*n* (%)	mtDNA content (IQR)	P
Breast cancer cell lines	Basal	5 (12.5%)	269 (149)	0.1†
	ERBB2	7 (17.5%)	620 (521)	
	Luminal	19 (47.5%)	518 (359)	
	Normal	9 (22.5%)	489 (142)	
Primary breast tumour specimens	Basal	65 (31.8%)	454 (287)	0.8†
	ERBB2	35 (17.2%)	566 (351)	
	Luminal A	56 (27.5%)	423 (224)	
	Luminal B	40 (19.6%)	514 (343)	
	Normal	8 (3.9%)	377 (286)	

Median mtDNA content [number of mtDNA molecules per cell] with interquartile range (IQR) for each group and corresponding probabilities (P value) for equal distribution using Kruskal-Wallis one-way analysis of variance (†).

mtDNA content and the mesenchymal characteristics

In vitro reduction of mtDNA content has been linked to changes in expression of the EMT-related genes *CDH1* [12, 14], *CDH2* [14], *ESRP1* [14], *FN1* [14], *MMP9* [14], *SNAI1* [14], *SNAI2* [14], *TGFB1* [12], *TGFBR1* [12], *TWIST1* [14] and *VIM* [12, 14]. To address whether a more mesenchymal phenotype is a physiological characteristic linked to low mtDNA content [12, 14], we analysed the relation between mtDNA content and the RNA expression levels of genes related to EMT. Expression data for the above mentioned genes were available for 40 of the 41 breast cancer cell lines and all 204 primary breast tumour specimens. Expression data of *TGFBR1* was excluded because the probe gave expression levels close to background noise. Correlation between gene expression levels and mtDNA content did not exceed a correlation coefficient ρ of 0.35, and we could not demonstrate statistical significance after correction for multiple testing (all $P > 0.027$, **Supplementary Table 1**) in the breast cancer cell lines. In our primary breast tumour specimens, correlation between mtDNA content and the expression of *ESRP1* ($\rho = 0.25$, $P < 0.001$), *SNAI1* ($\rho = 0.23$, $P < 0.001$) and *TGFB1* ($\rho = 0.18$, $P < 0.01$) was statistically significant after correction for multiple testing (**Supplementary Table 1**). To further explore the link between mtDNA content and EMT, we analysed the association between mtDNA content and the intrinsic subtypes of breast cancer,

which have been assigned with epithelial or mesenchymal characteristics [36, 38, 39]. Comparisons between the intrinsic subtypes for both the breast cancer cell lines as well as the primary tumour specimens did not show differences in mtDNA content among the subtypes (P > 0.05) (**Table 1**).

Association of tumour mtDNA content with established prognostic clinicopathological variables

In our patient cohort, we analysed tumour mtDNA content in relation to patient age at diagnosis, menopausal status, tumour size, histological grade, estrogen receptor status, progesterone receptor status and *ERBB2* amplification (**Table 2**). Because the currently used conventional histological grade (modified Bloom-Richardson) was not available for nearly 25% of our cohort – samples originated from multiple hospitals and from time periods when histological grading according to Bloom-Richardson was not common – we included a molecular grading system shown to be equivalent, the qRT-PCR genomic grade index (GGI) [40]. There were no statistically significant associations between the

Table 2 Association between clinicopathological variables and mtDNA content.

Variable	Group	*n* (%)	mtDNA content (IQR)	P
Age at diagnosis	≤ 40	21 (10.3%)	491 (532)	0.21†
	> 40-55	88 (43.1%)	433 (273)	
	> 55-70	64 (31.4%)	466 (259)	
	> 70	31 (15.2%)	546 (490)	
Menopausal status	Pre	99 (48.5%)	427 (323)	0.15#
	Post	105 (51.5%)	500 (280)	
Tumour size	≤ 2 cm	99 (48.5%)	421 (280)	0.019#
	> 2 cm	105 (51.5%)	514 (382)	
Genomic Grade Index	1	35 (17.2%)	440 (225)	0.028†
	2	59 (32.4%)	410 (260)	
	3	103 (50.5%)	523 (389)	
Estrogen receptor status	Negative	87 (42.7%)	483 (328)	0.12#
	Positive	115 (56.4%)	424 (290)	
Progesterone receptor status	Negative	97 (47.5%)	480 (335)	0.073#
	Positive	96 (47.1%)	413 (281)	
ERBB2 amplification	Negative	169 (82.8%)	454 (287)	0.46#
	Positive	29 (14.2%)	463 (385)	

Number of patients and corresponding median mtDNA content [number of mtDNA molecules per cell] with interquartile range (IQR) for each group and corresponding probabilities (P value) for either equal distribution using Mann-Whitney U test (#) or Kruskal-Wallis one-way analysis of variance (†). Due to missing values the numbers of samples per variable do not always add up to 204.

tumour mtDNA content and age at diagnosis, menopausal status, estrogen receptor status, progesterone receptor status or *ERBB2* amplification status (P > 0.05). However, tumours smaller than 2 cm had statistically significant lower mtDNA content (median 421 mtDNA molecules per cell) compared to tumours larger than 2 cm (median 514 mtDNA molecules per cell) (Mann-Whitney P = 0.019). In addition, tumour mtDNA content varied between the GGI groups (Kruskal-Wallis P = 0.028), with the highest mtDNA content in the GGI group representing poorly differentiated high grade 3 tumours (median 523 mtDNA molecules per cell). However, we did not observe a significant trend in tumour mtDNA content across the GGI groups (Cuzick's test for trend P = 0.066).

Distant metastasis-free survival and primary tumour mtDNA content

Finally, we studied in our patient cohort the prognostic value of tumour mtDNA content with respect to the length of distant metastasis-free survival. All included breast cancer patients presented as lymph node-negative and did not receive any (neo)adjuvant systemic treatment. The distribution of mtDNA content in our cohort was skewed and could not be normalized by transformation (Skewness and Kurtosis test P < 0.05). To assess

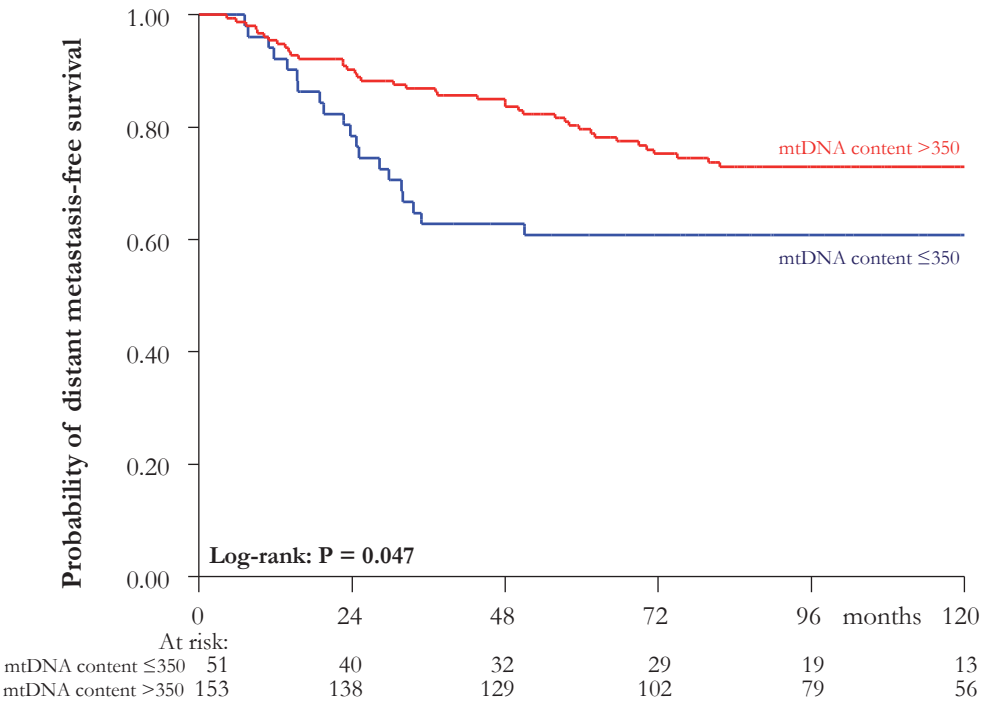


Figure 1 Kaplan-Meier curve showing probability of distant metastasis-free survival as a function of tumour mtDNA content of 204 patients (60 events). Numbers of patients at risk at 24 month time intervals are indicated.

tumour mtDNA content for the length of metastasis-free survival in our exploratory analysis, we first divided the cohort based on mtDNA content quartiles in four groups (Q1 – Q4). Because the patients within the first quarter Q1 presented a different rate

Table 3 Univariate and multivariable analyses for distant metastasis-free survival in lymph node-negative patients who did not receive any (neo)adjuvant systemic therapy.

			Univariate		Multivariable	
Variable	**Group**	***n* (%)**	**Hazard ratio (95% CI)**	**P**	**Hazard ratio (95% CI)**	**P**
Age	≤ 40	19 (10.2%)	1		1	
	> 40-55	81 (43.5%)	0.40 (*0.19-0.88*)	0.022	0.34 (*0.15-0.76*)	0.009
	> 55-70	59 (31.7%)	0.42 (*0.19-0.95*)	0.038	0.27 (*0.08-0.92*)	0.037
	> 70	27 (14.5%)	0.39 (*0.14-1.05*)	0.062	0.25 (*0.06-0.97*)	0.045
Menopausal status	Pre	91 (48.9%)	1		1	
	Post	95 (51.1%)	0.94 (*0.55-1.61*)	0.8	1.55 (*0.54-4.44*)	0.4
Tumour size	≤ 2 cm	87 (46.8%)	1		1	
	> 2 cm	99 (53.2%)	1.06 (*0.62-1.83*)	0.8	0.92 (*0.54-1.64*)	0.8
Genomic Grade Index	1	33 (17.7%)	1		1	
	2	56 (30.1%)	1.62 (*0.63-4.18*)	0.3	1.43 (*0.54-3.76*)	0.5
	3	97 (52.2%)	2.65 (*1.03-6.83*)	0.043	2.52 (*0.95-6.66*)	0.063
Progesterone receptor status	Negative	96 (51.6%)	1		1	
	Positive	90 (48.4%)	0.74 (*0.38-1.45*)	0.4	0.88 (*0.43-1.81*)	0.7
ERBB2 amplification	Negative	159 (85.5%)	1		1	
	Positive	27 (14.5%)	1.39 (*0.70-2.78*)	0.3	1.45 (*0.70-2.97*)	0.3
mtDNA content	≤350	48 (25.8%)	1		1	
	>350	138 (74.2%)	0.50 (*0.29-0.87*)	0.015	0.54 (*0.30-0.97*)	0.038

Number of patients and corresponding hazard ratio for distant metastasis-free survival with its 95% confidence intervals (CI) and corresponding probabilities for equal risk (P value) for each group. Analyses were stratified for estrogen receptor status and limited to the 186 patients (54 events) with no missing values.

of metastasis-free survival compared to the other quarters (Q2 – Q4) (**Supplementary Figure 1**), we divided the cohort in two patient groups of low mtDNA content (Q1 with mtDNA content ≤350 mtDNA molecules per cell) versus the rest (Q2 – Q4 with mtDNA content >350 mtDNA molecules per cell). To visualize the length of metastasis-free survival as a function of the levels of tumour mtDNA content (Q1 vs Q2 – Q4) we used the Kaplan-Meier survival analysis method (**Figure 1**). Patients in the low mtDNA content group Q1 showed a higher metastasis probability (log-rank P = 0.047). In univariate and multivariable Cox regression analysis including only the 186 patients with no missing values (**Table 3**), patients in the Q2 – Q4 mtDNA content group showed a longer distant metastasis-free survival compared to patients in the low mtDNA tumour content group (univariate: HR 0.50, 95% CI: 0.29-0.87, P = 0.015; multivariable: HR 0.54, 95% CI: 0.30-0.97, P = 0.038).

Discussion

Many contradictions about the physiological consequences of reduced mtDNA content exist in the literature. A critical reduction in mtDNA content compromises mitochondrial functioning with downstream effects. Subsequent changes in cellular processes such as aerobic respiration, calcium homeostasis or the intrinsic apoptotic pathway could in turn impact tumorigenic properties. Previous findings have pointed towards a link between low mtDNA content and breast cancer aggressiveness but the exact association remains uncertain. Here, to elucidate its potential as a prognostic marker, the putative relation between tumour mtDNA content and mesenchymal features or distant metastasis-free survival in breast cancer was explored. Using a quantitative PCR approach, the mtDNA content of 41 breast cancer cell lines and 204 primary breast tumour specimens was obtained. A correction for copy number variations of the nuclear-encoded reference locus (*HMBS*) minimized bias due to tumour-related genomic aberrations in the obtained number of mtDNA molecules per cell. Furthermore, the quantitative mtDNA target in our current assay lies outside of the common deletion region [41] and therefore it is likely that we measure a mixture of functional and dysfunctional mtDNA molecules.

Previous *in vitro* studies reported induction of EMT and stem-cell features upon depletion of mtDNA [12, 14, 19]. In the panel of breast cancer cell lines – homologous cell populations – no relation between mtDNA content and expression levels of genes involved in EMT could be demonstrated. In the cohort of primary breast tumour specimens – more heterogeneous cell populations – we find a positive but weak correlation ($\rho \leq 0.25$) between mtDNA content and *ESRP1*, *SNAI1* and *TGFB1*. In *in vitro* studies, a reduction in mtDNA content resulted in decreased ESRP1 protein levels

but, contradictory to our findings, increased *SNAI1* mRNA expression or TGFB protein expression. In addition, we could not demonstrate a difference in mtDNA content between the intrinsic subtypes in our cell line panel nor in the cohort of primary breast specimens. Mesenchymal properties have been attributed to the basal and normal-like subtypes, whereas the luminal subtypes are generally epithelial [36, 38, 39]. Accordingly, the evaluated EMT-related genes were commonly highly statistically significant related to each other and differentially expressed between the intrinsic subtypes within our cell line panel and the cohort of primary breast tumour specimens (**Supplementary Table 1**). Apart from a true lack of association between mtDNA and EMT-features in breast cancer, there are several other reasons which may explain the absence of this association in our data set. To understand the physiological effects of mtDNA content, previous studies suggesting a relation between mtDNA and EMT often used cell lines artificially depleted of mtDNA, termed rho0 clones [42]. The endogenous mtDNA content of the cell lines and primary tumour specimens in our study is a few hundred molecules per cell, which is still orders of magnitude higher than of the rho0 clones. Since the extent of mtDNA reduction is of importance in gaining tumorigenic properties, as demonstrated in glioblastoma models [43], perhaps the mtDNA content in our data set is not at the critically low level necessary to induce a transition towards a mesenchymal phenotype. Alternatively, low mtDNA levels may be important during the process of EMT, but might be restored to normal after the transition is accomplished. Despite the unknown exact reason, we conclude that in our data set mtDNA content is not related to the molecular features connected to a mesenchymal-like phenotype.

To address a possible relation between mtDNA content and aggressive behaviour *in vivo*, we analysed primary breast tumour mtDNA content and prognosis in a cohort of 204 breast cancer patients. Notably, the mtDNA content in our primary tumour specimens is only an estimate, representing not only a heterogeneous tumour cell population but also non-neoplastic cells incorporated in the tumour specimen. However, because no evidence for a relation between mtDNA content and tumour infiltrating lymphocytes [44] was observed (**Supplementary Table 2**) and stromal content was minimized (Materials and Methods), we estimate the contribution of non-neoplastic cells to the final mtDNA content to be minimal. A few associations between mtDNA content and clinicopathological variables have been reported in previous studies, albeit never consistently [24-29]. These studies included either a low number of study participants or heterogeneous groups regarding treatment regimen or disease stage, making interpretation difficult. In this study, we included a population of lymph node-negative primary breast cancer patients who did not receive any (neo)adjuvant systemic treatment. In this patient group, lower mtDNA content was observed in tumours smaller than 2 cm across compared to tumours larger than 2 cm across. Previous studies could not demonstrate

such a difference [27, 28] or reported lower mtDNA content in tumours over 5 cm across compared to smaller tumours [26]. Larger tumours presumably underwent more cell divisions potentially resulting in additional replication-induced mtDNA damage [45], which in turn might require additional compensatory mtDNA molecules to maintain proper mitochondrial functioning. It is also plausible that a hypoxic environment in larger tumours reduces mtDNA content as suggested previously [26]. However, in our primary tumour cohort no relation was observed between mtDNA content and hypoxia-related gene expression [46] as surrogate for the hypoxic state of the tumour (**Supplementary Table 2**). In addition, our results show a relation between mtDNA content and GGI, a gene expression-based identifier of the histological grade of tumours [40], with the highest grade representing poorly differentiated tumours showing higher mtDNA content. However, we could not demonstrate a conventional significant trend between mtDNA content and GGI. The relation between mtDNA content and histological grade has been reported before [26]. An increase in mtDNA content occurs in early S-phase of the cell cycle [47], and we attribute the relation between grade 3 tumours and higher mtDNA content to the high-proliferative nature of these higher grade tumours. Nevertheless, we note that the median difference in mtDNA content for both tumour size and GGI is only 20%, making a substantial biological consequence of these associations less likely.

Importantly, our cohort is highly suitable to study the prognostic value of mtDNA content for distant metastasis-free survival because all included patients presented with lymph node-negative disease and did not receive any (neo)adjuvant systemic treatment. The size of our cohort did not allow for separate analyses for estrogen receptor-negative and -positive tumours, which show different proportionality over time (test of proportional hazards assumption $P = 0.016$). Therefore, stratification for estrogen receptor status was applied in all proportional hazard analyses. After adjustment for established prognostic clinicopathological variables, we observed a prognostic effect for mtDNA content. The patients with the 25% lowest mtDNA content (≤350 mtDNA molecules per cell) showed a significant unfavourable prognosis with shorter time to metastasis compared to patients with higher mtDNA content. One previous study reported on low tumour mtDNA content corresponding to a higher risk of death [26]. However in that study, no clear information was provided about treatments administered, disease stage at diagnosis and other clinical variables included in their statistical analysis. This renders interpretation and comparison with that previous study difficult. Interestingly, low mtDNA content predicted for a favourable response to anthracycline treatment in a small patient cohort [24]. It is plausible that cells with low mtDNA content are susceptible to such regimen, because damage of mtDNA in cells containing fewer mtDNA molecules can affect mitochondrial functionality more effectively. In our cohort we could study the prognostic value of mtDNA content independent of treatment regimen.

To conclude, we demonstrate a link between particularly low mtDNA content and metastatic potential in breast cancer, which appears to be independent of the mesenchymal phenotype. Low primary tumour mtDNA content potentially identifies patients with unfavourable prognosis but at the same time might predict therapeutic efficacy of DNA-damaging treatment regimen in this group. Larger cohorts of uniformly treated patients are necessary to validate these results and to further unravel the clinical relevance of mtDNA content determination in cancer.

Materials and methods

Study cohort and sampling

We employed a panel of 42 breast cancer cell lines (including BT20, BT474, BT483, BT549, CAMA1, DU4475, EVSAT, HCC1937, Hs578T, MCF7, MDAMB134VI, MDAMB157, MDAMB175VII, MDAMB231, MDAMB330, MDAMB361, MDAMB415, MDAMB435s, MDAMB436, MDAMB453, MDAMB468, MPE600, OCUBF, OCUBM, SKBR3, SKBR5, SKBR7, SUM102PT, SUM1315MO2, SUM149PT, SUM159PT, SUM185PE, SUM190PT, SUM225CWN, SUM229PE, SUM44PE, SUM52PE, T47D, UACC812, UACC893, ZR751 and ZR7530 [36, 44]). In addition, DNA extracts from fresh frozen primary breast tumour specimens from an earlier study [37] were selected from our bio-bank at the Erasmus MC. The study was approved by the medical ethics committee of the Erasmus MC (MEC 02.953) and conducted in accordance to the code of conduct of Federation of Medical Scientific Societies in the Netherlands. Whenever possible, we adhered to the Reporting Recommendations for Tumour Marker Prognostic Studies (REMARK) [48]. Patient selection criteria have been described before [49] and include lymph node-negative primary breast cancer with local treatment but no systemic (neo)adjuvant therapies. Our selection (**Supplementary Figure 2**) was based on availability of genotypic data and gene expression data from the primary tumours (n = 337) and availability of uniformly extracted DNA (see below) (n = 250). Next, specimens with a tumour cell percentage below 50% were excluded to minimize skewed values due to stromal cell contamination (n = 38). In addition, five patient samples were ineligible in retrospect and excluded. Thus, mtDNA content was examined in a total of 207 patients. Patients' follow-up involved examinations every 3 months for the first two 2 years, every 6 months for years 3–5, and every 12 months from year 5 onwards. Estrogen receptor and progesterone receptor status were determined as described before [50]. Evaluation of *ERBB2* amplification via RNA expression levels and the qRT-PCR Genomic Grade Index were determined as described before respectively [51] and [40].

DNA extraction

DNA was extracted from cultured cell lines using the DNeasy Blood & Tissue kit (*Qiagen, Venlo, the Netherlands*) according the suppliers' protocol. We selected the DNA which was previously extracted from cryostat sections of the primary tumour tissues based on uniformity in extraction procedure (QIAamp DNA mini kit (*Qiagen*) as described before [37]). DNA extracts were quantified using the Qubit dsDNA HS assay kit (*Life Technologies, Carlsbad, United States of America*) and all samples were diluted to a concentration of 0.2 ng/μL DNA prior to mtDNA content analysis.

Copy number analysis

Copy number variation of the nuclear encoded *HMBS* gene – which served as a reference to obtain mtDNA content – was obtained from our previously described microarray data (Gene Expression Omnibus database accession numbers GSE10099 [37] and GSE41308 [52]). The breast cancer cell lines were genotyped on the Genome Wide Human SNP Array 6.0 (*Affymetrix, Santa Clara, United States of America*), the primary tumour specimens on the GeneChip Human Mapping 100K SNP Array (*Affymetrix*).

mtDNA content

Mitochondrial DNA content was determined in duplicate runs using a multiplex quantitative PCR targeting the nuclear *HMBS* gene (chr11q23.2-qter) and the mitochondrial *MT-TL1* (chrMT 3212–3319). Primers targeting the nuclear encoded *HMBS* gene (forward 5'-TGAGGCGGATGCAGATAC-3' and reverse 5'- CCCACCCACGGTAGTAATTC-3' (*Life technologies*)) yielded a 201 bp amplicon quantitatively detected using a CY5 labelled probe (5'-[CY5]TATCAGCCAAGCCTCCGAAC[BHQ2]-3' (*Sigma Aldrich, St. Louis, United States of America*)). Primers targeting the mitochondrial encoded *MT-TL1* (forward 5'-CACCCAAGAACAGGGTTTGT-3' and reverse 5'-TGGCCATGGGTATGTTGTTA-3' (*Life Technologies*)) yielded a 108 bp amplicon quantitatively detected using a HEX labelled probe (5'-[HEX] TTACCGGGCTCTGCCATCT[BHQ1]-3', (*Sigma Aldrich*)) [53]. Reactions included 1x Absolute QPCR Mix containing SYBR Green and ROX (*AB-1163 Life Technologies*) in the presence of 100 nM mtDNA primers, 360 nM nDNA primers and 100 nM probes. The 45-cycle PCR was carried out at a 62°C annealing temperature and probe fluorescence was monitored using ROX, HEX, CY5 and FAM filters on Mx3000P or Mx3005P qPCR systems (*Agilent Technologies, Waldbronn, Germany*). Quantification cycle values (Cq [dRN]) were obtained using the adaptive baseline approach (MxPro v4.10) up to cycle 35 with fixed fluorescence thresholds at 0.004 dRn. Performance of singleplex PCR and multiplex PCR runs was comparable (**Supplementary Table 3**). Performance

of the assay at variable ratios of artificial *HMBS* (289 bp linear: ACA GAC GGG GTC CTT TCA TTC GAG GCT GGG CTG AGG CGG ATG CAG ATA CGG CCC CTT TGG GAA GAC ACG TTC CAC TTT TGA TTC ATA GGA GAG AGT ATC AGC CAA GCC TCC GAA CTG CAC ACA AAC GTC TTA GAA GTG CGC CTT CTT TTT GTG TTA TAG TGG TCT CCC AGC CAC AGC CAA CGC TCC AAG TCC CCA GCT GTG ACA CAC CTA CTG AAT TAC TAC CGT GGG TGG GAG GCC GCC GTG GGC CTT TCC ATT ACG AGC CTG CTT GCC GAG CCC TGG GCT TGT GCA C) and artificial *MT-TL1* (180 bp cloned in circular 2374 bp pMA-T vector: TAT CAT CTC AAC TTA GTA TTA TAC CCA CAC CCA CCC AAG AAC AGG GTT TGT TAA GAT GGC AGA GCC CGG TAA TCG CAT AAA ACT TAA AAC TTT ACA GTC AGA GGT TCA ATT CCT CTT CTT AAC AAC ATA CCC ATG GCC AAC CTC CTA CTC CTC ATT GTA CCC ATT CTA ATC GCA ATG GCA) was linear (**Supplementary Table 4**). DNA input of the breast cancer cell lines and the primary breast tumour specimens was standardized for 1 ng DNA per reaction. A calibration curve containing a pool of DNA isolates from independent fresh frozen tumours was taken along as internal control to monitor inter assay variation. Obtained Cq values were used to calculate the ratio of mitochondrial DNA opposed to nuclear DNA by the relative quantitation method (2^ΔCq [54]). Multiplying this ratio by the copy number of *HMBS* (obtained as described above) resulted in the number of mtDNA molecules per cell as mtDNA content.

Gene expression analysis

Gene expression data of the cell lines was obtained from our previously described triplicate microarray data (Gene Expression Omnibus database accession number GSE41313 [52]) on the Human Genome HT_HG-U133_Plus_PM GeneChip 96-well arrays (*Affymetrix*). Data of all breast cancer cell lines were available with the exception of SUM225CWN. Gene expression data of the primary breast tumour specimens was obtained from our previously described microarray data (Gene Expression Omnibus database accession number GSE2034 [50] and GSE5327 [55]) on the Human Genome HG-U133a GeneChip 96-well arrays (*Affymetrix*). Subtype classification was based on expression of the intrinsic gene set defined by Perou et al [56]. Cell line DU4475 could not be classified to a subtype group and was therefore excluded from the intrinsic subtype analysis. For individual genes, levels based on log2 transformed distances to the geometric mean for each probe set were obtained for probe IDs 201131_s_at (*CDH1*), 203440_at (*CDH2*), 219121_s_at (*ESRP1*), 210495_x_at (*FN1*), 203936_s_at (*MMP9*), 219480_at (*SNAI1*), 213139_at (*SNAI2*), 203085_s_at (*TGFB1*), 213943_at (*TWIST1*) and 201426_s_at (*VIM*). The tumour infiltrating lymphocyte classification as low TIL and high TIL was

based on the immune signature probe-set by Massink et al [44]. Classification as low hypoxia response and high hypoxia response was based on the expression of the hypoxia-related gene signature as described before by Chi et al [57].

Statistical analyses

All analyses included the average mtDNA content obtained from the duplicate analysis for each individual sample. Data distribution was tested using Skewness-Kurtosis tests for normality. Numerical correlations between RNA expression levels and mtDNA content were investigated using the Spearman rank correlation and corrected for multiple testing using the false discovery rate controlling procedure [58]. Categorical comparisons of the intrinsic subtypes or grouped clinical variables and mtDNA content were employed using either Mann-Whitney U-tests (two groups) or Kruskal-Wallis one-way analysis of variance (multiple groups). When appropriate, we performed Cuzick's test for trend across ordered categorical variables. Kaplan-Meier survival plots and log-rank tests were used to assess the differences in time to distant metastasis between mtDNA content groups. Proportional hazard analyses for distant metastasis-free survival were performed using Cox proportional-hazards regression methods. We stratified for estrogen receptor status and censored for 10 years clinical follow-up (most patients are redirected to their general practitioner at that point in time) to maintain proportionality (test of proportional hazards assumption using the Schoenfeld residuals $P > 0.05$). Univariate analysis was done on the individual clinicopathological variables, multivariable analysis included all clinicopathological variables and mtDNA content. All statistical tests were two-sided, and P values smaller than 0.05 were considered as statistically significant. Clinical variables were statistically analysed in Stata version 13.1 (*StataCorp LP, College Station, United States of America*). Other analyses were performed using Spotfire 7.0.0 (*TIBCO, Palo Alto, United States of America*).

Supplementary data

Supplementary data for this article are available online at Oncotarget (http://www.oncotarget.com).

References

1. Wallace, D.C., *Mitochondria and cancer.* Nat Rev Cancer, 2012. **12**(10): p. 685-698.
2. A Dickinson, K.Y.Y., J Donoghue, M J Baker, R DW Kelly, M McKenzie, T G Johns and J C St. John, *The regulation of mitochondrial DNA copy number in glioblastoma cells.* Cell Death and Differentiation, 2013. **20**: p. 1644–1653.
3. Facucho-Oliveira, J.M., et al., *Mitochondrial DNA replication during differentiation of murine embryonic stem cells.* J Cell Sci, 2007. **120**(Pt 22): p. 4025-4034.
4. Gahan, M.E., et al., *Quantification of mitochondrial DNA in peripheral blood mononuclear cells and subcutaneous fat using real-time polymerase chain reaction.* Journal of Clinical Virology, 2001. **22**(3): p. 241-247.
5. Miller, F.J., et al., *Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age.* Nucleic Acids Research, 2003. **31**(11) e.61.
6. Mansouri, A., et al., *An alcoholic binge causes massive degradation of hepatic mitochondrial DNA in mice.* Gastroenterology, 1999. **117**(1): p. 181-190.
7. Casula, M., et al., *Infection with HIV-1 induces a decrease in mtDNA.* J Infect Dis, 2005. **191**(9): p. 1468-1471.
8. Moraes, C.T., et al., *mtDNA depletion with variable tissue expression: a novel genetic abnormality in mitochondrial diseases.* Am J Hum Genet, 1991. **48**(3): p. 492-501.
9. El-Hattab, A.W. and F. Scaglia, *Mitochondrial DNA depletion syndromes: review and updates of genetic basis, manifestations, and therapeutic options.* Neurotherapeutics, 2013. **10**(2): p. 186-198.
10. Kulawiec, M., et al., *Tumorigenic transformation of human breast epithelial cells induced by mitochondrial DNA depletion.* Cancer Biol Ther, 2008. **7**(11): p. 1732-1743.
11. Singh, K.K., et al., *Inter-genomic cross talk between mitochondria and the nucleus plays an important role in tumorigenesis.* Gene, 2005. **354**: p. 140-146.
12. Naito, A., et al., *Progressive tumor features accompany epithelial-mesenchymal transition induced in mitochondrial DNA-depleted cells.* Cancer Sci, 2008. **99**(8): p. 1584-1588.
13. Moro, L., et al., *Mitochondrial DNA depletion in prostate epithelial cells promotes anoikis resistance and invasion through activation of PI3K/Akt2.* Cell Death and Differentiation, 2009. **16**(4): p. 571-583.
14. Guha, M., et al., *Mitochondrial retrograde signaling induces epithelial-mesenchymal transition and generates breast cancer stem cells.* Oncogene, 2014. **33**(45): p. 5238-5250.
15. Pelicano, H., et al., *Mitochondrial respiration defects in cancer cells cause activation of Akt survival pathway through a redox-mediated mechanism.* Journal of Cell Biology, 2006. **175**(6): p. 913-923.
16. Yu, M., et al., *Depletion of mitochondrial DNA by ethidium bromide treatment inhibits the proliferation and tumorigenesis of T47D human breast cancer cells.* Toxicol Lett, 2007. **170**(1): p. 83-93.
17. Cavalli, L.R., M. VarellaGarcia, and B.C. Liang, *Diminished tumorigenic phenotype after depletion of mitochondrial DNA.* Cell Growth & Differentiation, 1997. **8**(11): p. 1189-1198.
18. Morais, R., et al., *Tumor-forming ability in athymic nude mice of human cell lines devoid of mitochondrial DNA.* Cancer Research, 1994. **54**(14): p. 3889-3896.
19. Magda, D., et al., *mtDNA depletion confers specific gene expression profiles in human cells grown in culture and in xenograft.* BMC Genomics, 2008. **9**: 521.
20. Tan, A.S., et al., *Mitochondrial Genome Acquisition Restores Respiratory Function and Tumorigenic Potential of Cancer Cells without Mitochondrial DNA.* Cell Metabolism, 2015. **21**(1): p. 81-94.
21. Guo, J.H., et al., *Frequent Truncating Mutation of TFAM Induces Mitochondrial DNA Depletion and Apoptotic Resistance in Microsatellite-Unstable Colorectal Cancer.* Cancer Research, 2011. **71**(8): p. 2978-2987.
22. Yu, M., *Generation, function and diagnostic value of mitochondrial DNA copy number alterations in human cancers.* Life Sciences, 2011. **89**(3-4): p. 65-71.

23. Cook, C.C. and M. Higuchi, *The awakening of an advanced malignant cancer: an insult to the mitochondrial genome.* Biochim Biophys Acta, 2012. **1820**(5): p. 652-662.
24. Hsu, C.W., et al., *Mitochondrial DNA Content as a Potential Marker to Predict Response to Anthracycline in Breast Cancer Patients.* Breast Journal, 2010. **16**(3): p. 264-270.
25. Yu, M., et al., *Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients.* Iubmb Life, 2007. **59**(7): p. 450-457.
26. Bai, R.K., et al., *Mitochondrial DNA content varies with pathological characteristics of breast cancer.* J Oncol, 2011. **2011**: 496189.
27. Tseng, L.M., et al., *Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer.* Genes Chromosomes Cancer, 2006. **45**(7): p. 629-638.
28. Fan, A.X., et al., *Mitochondrial DNA content in paired normal and cancerous breast tissue samples from patients with breast cancer.* J Cancer Res Clin Oncol, 2009. **135**(8): p. 983-989.
29. Barekati, Z., et al., *Methylation profile of TP53 regulatory pathway and mtDNA alterations in breast cancer patients lacking TP53 mutations.* Hum Mol Genet, 2010. **19**(15): p. 2936-2946.
30. Mambo, E., et al., *Tumor-specific changes in mtDNA content in human cancer.* Int J Cancer, 2005. **116**(6): p. 920-924.
31. McMahon, S. and T. LaFramboise, *Mutational patterns in the breast cancer mitochondrial genome, with clinical correlates.* Carcinogenesis, 2014. **35**(5): p. 1046-1054.
32. Thiery, J.P., *Epithelial-mesenchymal transitions in tumour progression.* Nat Rev Cancer, 2002. **2**(6): p. 442-454.
33. Naito, A., et al., *Induction of acquired resistance to antiestrogen by reversible mitochondrial DNA depletion in breast cancer cell line.* Int J Cancer, 2008. **122**(7): p. 1506-1511.
34. Yu, M., et al., *Mitochondrial DNA depletion promotes impaired oxidative status and adaptive resistance to apoptosis in T47D breast cancer cells.* Eur J Cancer Prev, 2009. **18**(6): p. 445-457.
35. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-10874.
36. Hollestelle, A., et al., *Loss of E-cadherin is not a necessity for epithelial to mesenchymal transition in human breast cancer.* Breast Cancer Res Treat, 2013. **138**(1): p. 47-57.
37. Zhang, Y., et al., *Copy Number Alterations that Predict Metastatic Capability of Human Breast Cancer.* Cancer Research, 2009. **69**(9): p. 3795-3801.
38. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): R68.
39. Sarrio, D., et al., *Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype.* Cancer Research, 2008. **68**(4): p. 989-997.
40. Toussaint, J., et al., *Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues.* BMC Genomics, 2009. **10**: 424.
41. Cortopassi, G.A. and N. Arnheim, *Detection of a Specific Mitochondrial-DNA Deletion in Tissues of Older Humans.* Nucleic Acids Research, 1990. **18**(23): p. 6927-6933.
42. King, M.P. and G. Attardi, *Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation.* Science, 1989. **246**(4929): p. 500-503.
43. Dickinson, A., et al., *The regulation of mitochondrial DNA copy number in glioblastoma cells.* Cell Death Differ, 2013. **20**(12): p. 1644-1653.
44. Massink, M.P., et al., *Genomic profiling of CHEK2*1100delC-mutated breast carcinomas.* BMC Cancer, 2015. **15**: 877.
45. Ju, Y.S., et al., *Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer.* Elife, 2014. **3**: e02935.
46. Chi, J.T., et al., *Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers.* PLoS Med, 2006. **3**(3): e47.
47. Trinei, M., et al., *Mitochondrial DNA copy number is regulated by cellular proliferation: A role for Ras and p66 (Shc).* Biochimica Et Biophysica Acta-Bioenergetics, 2006. **1757**(5-6): p. 624-630.

48. McShane, L.M., et al., *Reporting recommendations for tumor marker prognostic studies (REMARK).* J Natl Cancer Inst, 2005. **97**(16): p. 1180-1184.
49. Smid, M., et al., *Subtypes of breast cancer show preferential site of relapse.* Cancer Research, 2008. **68**(9): p. 3108-3114.
50. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-679.
51. van Agthoven, T., et al., *Relevance of breast cancer antiestrogen resistance genes in human breast cancer progression and tamoxifen resistance.* Journal of Clinical Oncology, 2009. **27**(4): p. 542-549.
52. Riaz, M., et al., *miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs.* Breast Cancer Res, 2013. **15**(2): R33.
53. Bai, R.K. and L.J.C. Wong, *Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease.* Journal of Molecular Diagnostics, 2005. **7**(5): p. 613-622.
54. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method.* Methods, 2001. **25**(4): p. 402-408.
55. Yu, J.X., et al., *Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.* BMC Cancer, 2007. **7**: 182.
56. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-752.
57. Chi, J.T., et al., *Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers.* Plos Medicine, 2006. **3**(3): e47.
58. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate - a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

CHAPTER 3


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Asp
Ser
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Leu

Low tumour mitochondrial DNA content is associated with better outcome in breast cancer patients receiving anthracycline-based chemotherapy

Marjolein J.A. Weerts | Antoinette Hollestelle | Anieta M. Sieuwerts | John A. Foekens | Stefan Sleijfer | John W.M. Martens

Clinical Cancer Research 2017; 23(16):4735-4743

Abstract

In this study, we aimed to explore whether low levels of mitochondrial DNA (mtDNA) content in the primary tumour could predict better outcome for breast cancer patients receiving anthracycline-based therapies. We hypothesized that tumour cells with low mtDNA content are more susceptible to mitochondrial damage induced by anthracyclines, and thus are more susceptible to anthracycline treatment. We measured mtDNA content by a quantitative PCR approach in 295 primary breast tumour specimens originating from two well-defined cohorts: 174 lymph node-positive patients who received adjuvant chemotherapy and 121 patients with advanced disease who received chemotherapy as first-line palliative treatment. The chemotherapy regimens given were either anthracycline-based (FAC/FEC) or methotrexate-based (CMF). In both the adjuvant and advanced setting, we observed increased benefit for patients with low mtDNA content in their primary tumour, but only when treated with FAC/FEC. In multivariable Cox regression analysis for respectively distant metastasis-free survival and progression-free survival, the hazard ratio for the FAC/FEC treated mtDNA low group in the adjuvant setting was 0.46 (95% confidence interval (CI) 0.24-0.89, P = 0.020) and in the advanced setting 0.49 (95% CI 0.27-0.90, P = 0.022) compared to the FAC/FEC treated mtDNA high group. We did not observe these associations in the patients treated with CMF. In our two study cohorts, breast cancer patients with low mtDNA content in their primary tumour have better outcome from anthracycline-containing chemotherapy. The frequently observed decrease in mtDNA content in primary breast tumours may be exploited by guiding chemotherapeutic regimen decision-making.

Introduction

Mitochondria are cellular organelles involved in multiple cellular processes, but best known for efficient ATP generation through oxidative phosphorylation. Mitochondria contain their own genomic entity termed mitochondrial DNA (mtDNA) encoding proteins essential for the oxidative phosphorylation system, and ribosomal RNAs and transfer RNAs functioning in the mitochondrial translation apparatus. Multiple mtDNA molecules can reside within a single mitochondrion and multiple mitochondria can reside within a single cell [1-3], making the total number of mtDNA molecules per cell (mtDNA content) variable. In general, the mtDNA content per cell is dependent on the tissue's energy demands [4].

In tumours, the mtDNA content is often changed compared to non-neoplastic adjacent tissue [5]. For breast cancer specifically, there is a decline in mtDNA content: approximately three-quarter of primary breast tumour specimens have a decreased mtDNA content when compared to their nearby normal mammary epithelium [5-13]. We recently reported an association of worse 10-year distant metastasis-free survival for node-negative primary breast cancer patients who did not receive any (neo-)adjuvant systemic treatment with low mtDNA content in their primary tumours, showing the impact of mtDNA content on tumour aggressiveness [14]. However, how these findings influence response to systemic therapy in breast cancer patients is unknown.

The anthracyclines doxorubicin and epirubicin are currently the most frequently used agents in breast cancer treatment. However, despite multiple efforts to find predictors for anthracycline sensitivity, up to date no evidence-based biomarkers are applied clinically in neither early nor metastatic breast cancer. Several markers have been postulated to predict benefit from adjuvant anthracycline-based chemotherapy, including *TOP2A* gene amplification or protein expression, *ERBB2* (HER2) amplification, *TOP2A* and *ERBB2* co-amplification, chromosome 17 polysomy (CEP17), TIMP1 protein expression, FOXP3 protein expression or TP53 protein expression, but none of them have been recommended for clinical use [15]. Anthracyclines induce severe oxidative stress [16] and are known to accumulate in mitochondria, where they can intercalate mtDNA [17] and damage mtDNA [18]. In *in vitro* model systems, reduced mtDNA content increases sensitivity to doxorubicin [13, 19]. We hypothesize that tumour cells with low mtDNA content are more susceptible to mitochondrial damage induced by anthracyclines than cells with high mtDNA content, and thus are more susceptible to anthracycline treatment. Before the introduction of anthracycline-based chemotherapy, methotrexate-based regimens were most often applied in breast cancer [20]. The working mechanism of this compound is not directly at the DNA level, but it is an anti-metabolite, ultimately leading to inhibition of DNA synthesis. Since methotrexate induces only low levels of oxidative stress [16], we

do not expect differences in its efficacy in tumour cells with either low or high mtDNA content.

In this retrospective study, we aimed to explore if low levels of mtDNA content in the primary tumour can predict better outcome for patients receiving anthracycline-based therapies, but not for patients receiving methotrexate-based therapies.

Results

mtDNA content in primary breast tumours

In total, we analysed primary tumour DNA from 295 breast cancer patients derived from two cohorts: node-positive patients receiving either FAC/FEC or CMF adjuvant chemotherapy treatment (adjuvant cohort), and patients receiving either FAC/FEC or CMF first-line palliative chemotherapy treatment for their recurrent disease (advanced cohort). We obtained mtDNA content in these DNA extracts by multiplex real time quantitative PCR (qPCR) targeting a nuclear-encoded and a mitochondrial-encoded gene, and included a correction for sample-specific somatic copy number variation of the nuclear-encoded gene as obtained by qPCR.

In the adjuvant cohort of 174 patients, the median mtDNA content was 758 mtDNA molecules per cell (interquartile range (IQR) 506), in the advanced cohort of 121 patients the median mtDNA content was 694 mtDNA molecules per cell (IQR 402). Because mtDNA content was not normally distributed within both cohorts (both Shapiro-Wilk $P < 0.001$), we dichotomized each cohort based on the median mtDNA content, resulting in mtDNA low and mtDNA high groups. No differences were observed with respect to the chemotherapeutic regimens given between the mtDNA low and mtDNA high groups in both the adjuvant and the advanced cohorts (Fisher's exact $P = 0.4$ and $P = 0.5$, respectively) (**Supplementary Table 5**).

mtDNA content in association with clinicopathological variables

Next, we evaluated the association between mtDNA content and clinicopathological variables. In the adjuvant cohort, no statistically significant associations were observed between mtDNA content and age at diagnosis, menopausal status at diagnosis, nodal status (number of nodes positive), primary tumour size (T-stage), primary tumour Genomic Grade Index (GGI), estrogen receptor (ER), progesterone receptor (PR) and HER2 status (Fisher's exact $P > 0.05$) (**Table 1**). In the advanced cohort, no statistically significant associations were observed between mtDNA content and age at recurrence, menopausal status at recurrence, nodal status, primary tumour size, primary tumour GGI, PR status, HER2 status, dominant relapse site, disease-free interval and treatment

Table 1 Association between mtDNA content and clinicopathological variables of node-positive patients receiving either FAC/FEC or CMF adjuvant chemotherapy treatment (adjuvant cohort).

Characteristic	mtDNA low *n* (%)	mtDNA high *n* (%)	P
Age (at diagnosis)			
≤ 40	27 (31.0%)	19 (21.8%)	0.3
> 40-50	43 (49.4%)	53 (60.9%)	
> 50	17 (19.5%)	15 (17.2%)	
Menopausal status (at diagnosis)			
Pre	80 (92.0%)	74 (85.1%)	0.2
Post	7 (8.0%)	13 (14.9%)	
Tumour size			
T1 (≤ 2 cm)	24 (29.3%)	13 (15.3%)	0.1
T2 (2-5 cm)	43 (52.4%)	54 (63.5%)	
T3/4 (> 5 cm)	15 (18.3%)	18 (21.2%)	
unknown	*5*	*2*	
Nodal status			
1-3	53 (60.9%)	47 (54%)	0.4
> 3	34 (39.1%)	40 (46%)	
Grade (GGI)			
1	15 (19.7%)	11 (15.7%)	0.5
2	27 (35.5%)	32 (45.7%)	
3	34 (44.7%)	27 (38.6%)	
unknown	*11*	*17*	
ER status			
Negative	24 (27.6%)	24 (27.6%)	1
Positive	63 (72.4%)	63 (72.4%)	
PR status			
Negative	24 (29.6%)	27 (33.3%)	0.7
Positive	57 (70.4%)	54 (66.7%)	
unknown	*6*	*6*	
HER2 status			
Balanced	64 (82.1%)	59 (83.1%)	1
Amplified	14 (17.9%)	12 (16.9%)	
unknown	*9*	*16*	

with consolidation therapy (Fisher's exact P > 0.05) (**Table 2**), but ER-positive primary tumours were more prevalent in the mtDNA low group (Fisher's exact P = 0.018).

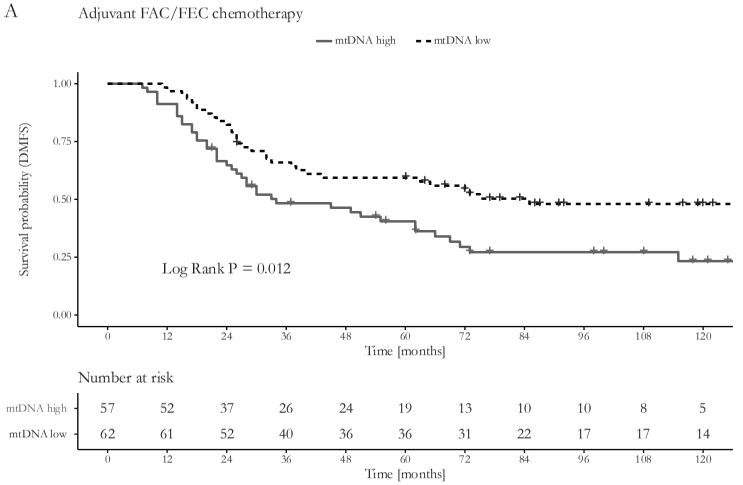
mtDNA content in association with patient outcome

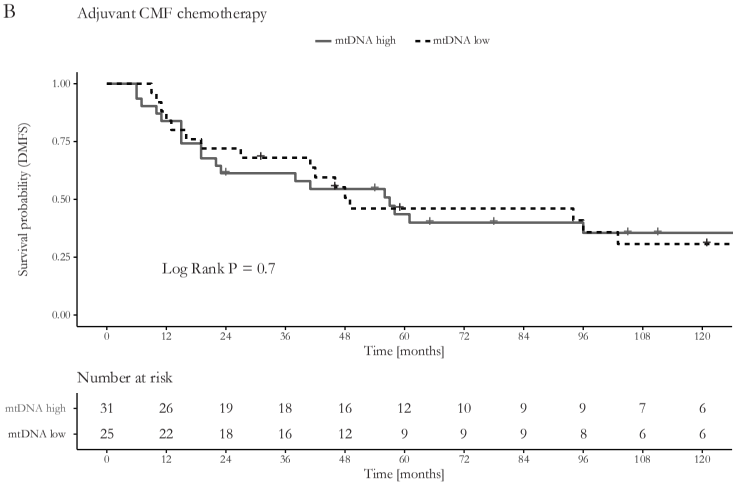
In the adjuvant cohort, patients treated with adjuvant FAC/FEC showed a significant longer distant metastasis-free survival (DMFS) when their primary tumour had low mtDNA content compared to patients with a high tumour mtDNA content (Log Rank P = 0.015) (**Figure 1A**). The median length of DMFS in the mtDNA low and mtDNA high groups were respectively 85 months (95% CI 40-NA) and 34 months (95% CI 26-69). The

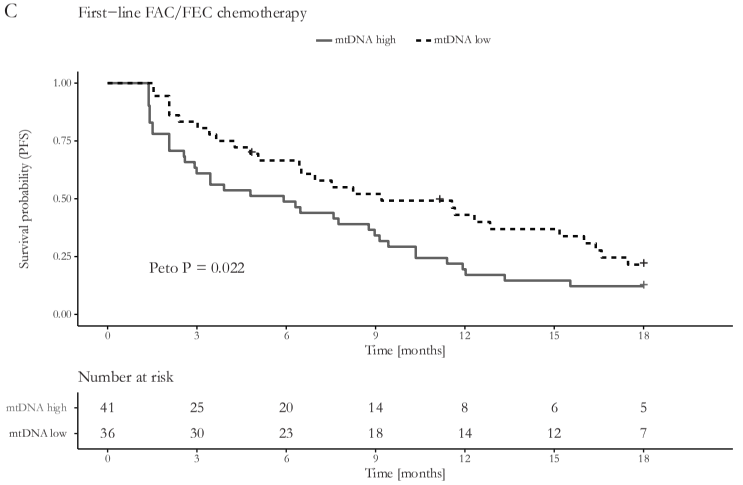
Table 2 Association between mtDNA content and clinicopathological of patients receiving either FAC/FEC or CMF first-line chemotherapy treatment for their recurrent disease (advanced cohort).

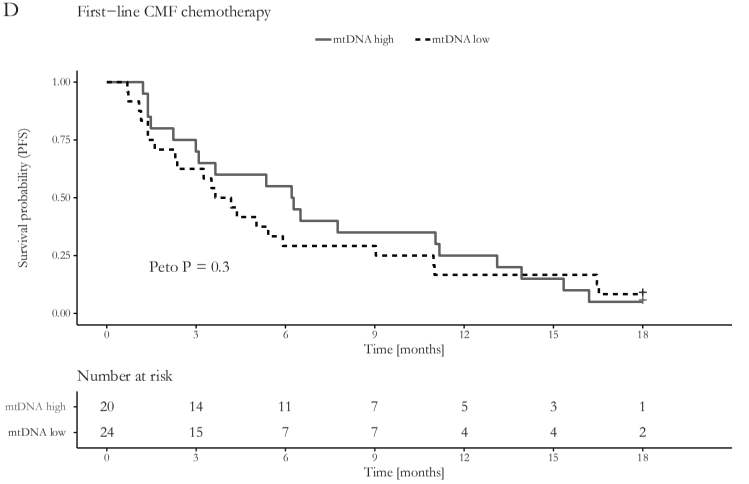
Characteristic	mtDNA low *n* (%)	mtDNA high *n* (%)	P
Age (at recurrence)			
≤ 45	14 (23.3%)	20 (32.8%)	0.1
> 45-55	25 (41.7%)	15 (24.6%)	
> 55	21 (35.0%)	26 (42.6%)	
Menopausal status (at recurrence)			
Pre	30 (50.0%)	31 (50.8%)	1
Post	30 (50.0%)	30 (49.2%)	
Tumour size			
T1 (≤ 2 cm)	17 (28.8%)	17 (28.3%)	0.8
T2 (2-5 cm)	32 (54.2%)	30 (50.0%)	
T3/4 (> 5 cm)	10 (16.9%)	13 (21.7%)	
unknown	*1*	*1*	
Nodal status			
0	17 (28.3%)	25 (41.0%)	0.3
1-3	15 (25.0%)	13 (21.3%)	
> 3	28 (46.7%)	23 (37.7%)	
Grade (GGI)			
1	8 (15.1%)	5 (8.5%)	0.3
2	14 (26.4%)	22 (37.3%)	
3	31 (58.5%)	32 (54.2%)	
unknown	*7*	*2*	
ER status			
Negative	26 (43.3%)	40 (65.6%)	0.018
Positive	34 (56.7%)	21 (34.4%)	
PR status			
Negative	38 (63.3%)	41 (67.2%)	0.7
Positive	22 (36.7%)	20 (32.8%)	
HER2 status			
Balanced	38 (71.7%)	46 (76.7%)	0.7
Amplified	15 (28.3%)	14 (23.3%)	
unknown	*7*	*1*	
Dominant relapse site			
Soft tissue	5 (8.3%)	6 (9.8%)	0.8
Bone	8 (13.3%)	11 (18.0%)	
Visceral	47 (78.3%)	44 (72.1%)	
Disease-free interval			
≤ 1 year	19 (31.7%)	26 (42.6%)	0.3
> 1-3 years	29 (48.3%)	28 (45.9%)	
> 3 years	12 (20.0%)	7 (11.5%)	
Hormonal consolidation therapy			
No	44 (74.6%)	47 (77.0%)	0.8
Yes	15 (25.4%)	14 (23.0%)	
unknown	*1*		

Figure 1 Kaplan-Meier survival curves as a function of mtDNA content for DMFS in patients receiving FAC/FEC (A) or CMF (B) adjuvant chemotherapy, and for PFS in patients receiving FAC/FEC (C) or CMF (D) first-line chemotherapy for their recurrent disease.


A
Adjuvant FAC/FEC chemotherapy
mtDNA high
mtDNA low
Survival probability (DMFS)
1.00
0.75
0.50
0.25
0.00
Log Rank P = 0.012
0 12 24 36 48 60 72 84 96 108 120
Time [months]
Number at risk
mtDNA high 57 52 37 26 24 19 13 10 10 8 5
mtDNA low 62 61 52 40 36 36 31 22 17 17 14
0 12 24 36 48 60 72 84 96 108 120
Time [months]


B
Adjuvant CMF chemotherapy
mtDNA high
mtDNA low
Survival probability (DMFS)
1.00
0.75
0.50
0.25
0.00
Log Rank P = 0.7
0 12 24 36 48 60 72 84 96 108 120
Time [months]
Number at risk
mtDNA high 31 26 19 18 16 12 10 9 9 7 6
mtDNA low 25 22 18 16 12 9 9 9 8 6 6
0 12 24 36 48 60 72 84 96 108 120
Time [months]


C
First−line FAC/FEC chemotherapy
mtDNA high
mtDNA low
Survival probability (PFS)
1.00
0.75
0.50
0.25
0.00
Peto P = 0.022
0 3 6 9 12 15 18
Time [months]
Number at risk
mtDNA high 41 25 20 14 8 6 5
mtDNA low 36 30 23 18 14 12 7
0 3 6 9 12 15 18
Time [months]


D
First−line CMF chemotherapy
mtDNA high
mtDNA low
Survival probability (PFS)
1.00
0.75
0.50
0.25
0.00
Peto P = 0.3
0 3 6 9 12 15 18
Time [months]
Number at risk
mtDNA high 20 14 11 7 5 3 1
mtDNA low 24 15 7 7 4 4 2
0 3 6 9 12 15 18
Time [months]

5-year DMFS in the mtDNA low and mtDNA high groups were respectively 59% (95% CI 48-73) and 41% (95% CI 30-57). In univariable Cox regression analysis for DMFS, the mtDNA low group compared to the mtDNA high group has a hazard ratio (HR) of 0.57 (95% CI 0.36-0.9, P = 0.017) (**Table 3**). Also in multivariable Cox regression analysis corrected for the base model including traditional prognostic factors, the patients in the low mtDNA content group had a longer DMFS compared to the mtDNA high group (HR 0.46, 95% CI 0.24-0.89, P = 0.020) (**Table 3**). However, also in the adjuvant cohort, patients treated with adjuvant CMF showed an equal distant metastasis probability in the mtDNA low and mtDNA high groups (Log Rank P = 0.7) (**Figure 1B**). The median length

Table 3 Univariable and multivariable Cox regression analysis for DMFS of node-positive patients receiving FAC/FEC adjuvant chemotherapy (adjuvant cohort).

		Univariable			Multivariable	
Characteristic	**events/*n***	**HR (95% CI)**	**P**	**events/*n***	**HR (95% CI)**	**P**
Age (at diagnosis)						
≤ 40	24/33	1		19/25	1	
> 40-50	39/66	0.67 (0.40-1.11)	0.1	23/44	0.34 (0.17-0.67)	0.002
> 50	10/19	0.48 (0.23-1.01)	0.054	10/17	0.78 (0.33-1.85)	0.6
Menopausal status (at diagnosis)						
Pre	66/108	1		46/78	1	
Post	7/10	0.96 (0.44-2.10)	0.9	6/8	0.93 (0.33-2.60)	0.9
Tumour size						
T1 (≤ 2 cm)	14/28	1		12/21	1	
T2 (2-5 cm)	43/68	1.70 (0.93-3.12)	0.09	31/53	1.15 (0.54-2.45)	0.7
T3/4 (> 5 cm)	14/19	2.16 (1.02-4.55)	0.043	9/12	1.67 (0.62-4.49)	0.3
Nodal status						
1-3	41/73	1			1	
> 3	32/45	1.48 (0.93-2.36)	0.1		1.39 (0.76-2.55)	0.3
Grade (GGI)						
1	7/16	1		6/13	1	
2	24/38	1.84 (0.79-4.28)	0.2	22/34	1.37 (0.50-3.74)	0.5
3	28/44	2.03 (0.88-4.64)	0.1	24/39	1.21 (0.44-3.34)	0.7
ER status						
Negative	22/34	1		17/26	1	
Positive	51/84	0.76 (0.46-1.26)	0.3	35/60	1.68 (0.55-5.13)	0.4
PR status						
Negative	22/32	1		20/30	1	
Positive	44/74	0.64 (0.38-1.07)	0.1	32/56	0.41 (0.13-1.25)	0.1
HER2 status						
Balanced	48/85	1		42/74	1	
Amplified	12/14	2.54 (1.33-4.85)	0.004	10/12	2.04 (0.92-4.54)	0.1
mtDNA						
High	39/56	1		26/38	1	
Low	34/62	0.57 (0.36-0.90)	0.017	26/48	0.46 (0.24-0.89)	0.020

In the multivariable model, analysis was limited to 86 patients (52 events) with no missing values.

of DMFS in the mtDNA low and mtDNA high groups were respectively 49 months (95% CI 41-NA) and 57 months (95% CI 22-NA). The 5-year DMFS in the mtDNA low and mtDNA high groups was respectively 46% (95% CI 30-71) and 44% (95% CI 29-66). In univariable Cox regression analysis for DMFS, the mtDNA low group compared to the mtDNA high group has a HR of 0.88 (95% CI 0.46-1.69, P = 0.7).

In the advanced cohort, in logistic regression analysis for overall response, patients treated with first-line FAC/FEC for their recurrent disease in the mtDNA low group had an odds ratio (OR) of 2.48 (95% CI 0.97-6.61, P = 0.06) compared to patients in the mtDNA high group (**Table 4**). These patients with a low primary tumour mtDNA content had a significant longer progression-free survival (PFS) compared to patients with high mtDNA content in their primary tumour (Peto P = 0.022) (**Figure 1C**). The median length of PFS in the mtDNA low and mtDNA high groups were respectively 9.20 months (95% CI 6.44-16.39) and 5.91 months (95% CI 2.99-9.13). In univariable Cox regression analysis for PFS, the mtDNA low group compared to the mtDNA high group has a HR of 0.6 (95% CI 0.36-1.0, P = 0.048) (**Table 5**). Also in multivariable Cox regression analysis corrected for the base model including also traditional predictive factors, the patients in the low mtDNA content group had a longer PFS compared to the mtDNA high group with a HR of 0.49 (95% CI 0.27-0.90, P = 0.022) (**Table 5**). For the patients treated with first-line CMF for their recurrent disease, in logistic regression analysis for overall response the mtDNA low group had an OR of 0.48 (95% CI 0.14-1.57, P = 0.2) (**Table 4**). No significant differences were observed between these two groups in their probability to disease progression (Peto P = 0.3) (**Figure 1D**), with a median PFS of 3.91 months (95% CI 2.37-9.03) and 6.24 months (95% CI 3.09-13.11) in the mtDNA low and mtDNA high groups respectively. In univariable Cox regression analysis for PFS, the HR of the mtDNA low group compared to the mtDNA high group was 1.14 (95% CI 0.61-2.13, P = 0.7).

Table 4 Overall response rate of patients receiving either FAC/FEC or CMF first-line chemotherapy for their recurrent disease (advanced cohort).

	mtDNA low ***n*** **(%)**	**mtDNA high** ***n*** **(%)**	**OR (95% CI)**	**P**
First-line FAC/FEC				
Non-responders	10 (27.8%)	20 (48.8%)	2.48 (0.97-6.61)	0.06
Responders	26 (72.2%)	21 (51.2%)		
First-line CMF				
Non-responders	14 (58.3%)	8 (40.0%)	0.48 (0.14-1.57)	0.2
Responders	10 (41.7%)	12 (60.0%)		

Table 5 Univariable and multivariable Cox regression analysis for PFS of patients receiving FAC/FEC first-line chemotherapy for their recurrent disease (advanced cohort).

		Univariable			Multivariable	
Characteristic	**events/*n***	**HR (95% CI)**	**P**	**events/*n***	**HR (95% CI)**	**P**
Age (at recurrence)						
≤ 45	19/24	1		17/21	1	
> 45-55	26/30	1.15 (0.64-2.09)	0.6	24/28	1.74 (0.77-3.93)	0.2
> 55	18/23	0.97 (0.51-1.85)	0.9	17/21	0.85 (0.25-2.88)	0.8
Menopausal status (at recurrence)						
Pre	37/44	1		33/39	1	
Post	26/33	0.85 (0.51-1.40)	0.5	25/31	0.90 (0.35-2.35)	0.8
Dominant relapse site						
Soft tissue	5/5	1		5/5	1	
Bone	8/10	0.13 (0.04-0.42)	<0.001	7/9	0.07 (0.02-0.28)	<0.001
Visceral	50/62	0.18 (0.06-0.48)	<0.001	46/56	0.17 (0.05-0.55)	0.003
Disease-free interval						
≤ 1 year	25/26	1		24/25	1	
> 1-3 years	28/37	0.73 (0.42-1.26)	0.3	24/32	1.06 (0.55-2.04)	0.9
> 3 years	10/14	0.65 (0.31-1.36)	0.3	10/13	0.85 (0.33-2.15)	0.7
ER status						
Negative	38/43	1		35/40	1	
Positive	25/34	0.58 (0.35-0.96)	0.036	23/30	1.21 (0.56-2.63)	0.6
PR status						
Negative	45/52	1		42/48	1	
Positive	18/25	0.69 (0.40-1.20)	0.2	16/22	0.85 (0.33-2.14)	0.7
HER2 status						
Balanced	44/55	1		44/54	1	
Amplified	14/16	1.31 (0.71-2.39)	0.4	14/16	1.07 (0.52-2.17)	0.9
Hormonal consolidation therapy						
No	51/53	1		46/48	1	
Yes	12/23	0.21 (0.11-0.40)	<0.001	12/22	0.16 (0.07-0.36)	<0.001
mtDNA						
High	36/41	1		35/40	1	
Low	27/36	0.60 (0.36-1.00)	0.048	23/30	0.49 (0.27-0.90)	0.022

In the multivariable model, analysis was limited to 70 patients (58 events) with no missing values.

Discussion

In this retrospective study, we measured mtDNA content in primary breast tumours and demonstrate that patients with low mtDNA content in their tumours have increased benefit from anthracycline-based chemotherapies (FAC/FEC), in both the adjuvant and advanced disease-setting. We hypothesized that tumour cells with low mtDNA content are more susceptible to mitochondrial damage induced by anthracyclines, and thus are more susceptible to anthracycline treatment. A previous study reported similar findings: for 27 patients receiving anthracycline-based adjuvant chemotherapy [13], the 5-year

disease-free survival (tumour recurrence) was 84% in the mtDNA low group opposed to 50% in the mtDNA high group, and this difference was not observed in 24 patients who did not receive adjuvant chemotherapy. However, these cohorts had only small numbers of patients, were heterogeneous regarding prognosis (i.e. lymph node status), and it was not reported if these patients received adjuvant hormonal treatment to improve outcome. In this study, we measured mtDNA content in nearly 300 primary breast tumours of two well-defined cohorts: lymph node-positive patients receiving only adjuvant chemotherapy as systemic treatment for their disease, and patients with disease recurrence receiving chemotherapy as first-line palliative treatment.

We measured mtDNA content using a multiplex qPCR approach where the mitochondrial-encoded locus lies outside of the common deletion region [21]. Also, we included a correction for copy number variation of the nuclear-encoded reference locus to minimize a bias in mtDNA content determination due to tumour-related local genomic aberrations. By including a minimal tumour cell percentage of 50% (the fraction of tumour cell nuclei within the cryosection) we aimed to minimize the contribution of non-tumour cells (i.e. stromal and immune cells) to the final mtDNA content. It is important to note that we evaluated the mtDNA content in the primary tumour specimens without comparing it to the tumour-adjacent normal mammary epithelium, and thus do not know whether the specimens with low mtDNA content have this due to variation of germline origin, or whether it is truly a somatic reduction in mtDNA copy number. In several other studies, mtDNA content in breast tumours has been compared with adjacent normal mammary tissue. In these studies, a somatic reduction of mtDNA was observed in ~70% of the cases [5-13], which makes it highly likely that those samples with a low mtDNA content in our dataset largely reflect the tumour-specific mtDNA content and is not due to germline origin.

Besides the stage in the disease trajectory, there are some differences between the two cohorts used in this study. The patient cohort receiving adjuvant chemotherapy was a younger cohort compared to the patients in the advanced cohort, with the majority of the patients being pre-menopausal. Also, the ER-positive cases were underrepresented in the advanced cohort compared to the general breast cancer population (approx. 75%), likely because ER-positive patients received (also) hormonal therapy (i.e. tamoxifen) and were thus excluded from our selection. In this advanced cohort, we observed an association between mtDNA content and ER status, with more ER-positive cases in the mtDNA content low group (**Table 2**). In both the adjuvant cohort of lymph node-positive patients, and in our previous study with lymph node-negative patients [14] or in other studies [7, 9, 12], no association between mtDNA content and ER status was observed. In addition, in the advanced cohort there were some differences between patients treated with first-line CMF or with first-line FAC/FEC: patients receiving CMF were older

(Fisher's exact P = 0.031, **Supplementary Table 7**) and received less often consolidation therapy (Fisher's exact P = 0.048, **Supplementary Table 7**). Also, similar to what we have done in the past for biomarker studies in the advanced setting, mtDNA content was measured in the primary tumour and not in the recurrent site. To our knowledge, it remains to be elucidated if the mtDNA content level in the recurrence site is similar to the primary tumour. Despite this, we did observe the hypothesized association between low mtDNA content and PFS in patients receiving first-line FAC/FEC chemotherapy, and this association was independent of established clinicopathological variables affecting outcome. We did not observe a statistically significant association with overall response (complete remission, partial remission or stable disease > 6 courses) within this group, but post-hoc power calculation with the current findings (72% response in mtDNA low, 51% response in mtDNA high, **Table 4**) indicates that our power is currently 45%. The objective response – which does not include patients with stable disease but only those with complete and partial remission or progressive disease and thus only takes into account the two extremes – in logistic regression analysis for the mtDNA low group compared to the mtDNA high group gives an OR of 4.11 (95% CI 1.27-15.17, P = 0.02) in the FAC/FEC group, and an OR of 0.6 (95% CI 0.12-3.00, P = 0.23) in the CMF group, however these patient numbers are very small (**Supplementary Table 3**). A total number of 166 patients will be necessary to elucidate with 80% power if there is a difference in overall response to first-line FAC/FEC in the mtDNA low and mtDNA high group.

Several markers have been described to predict benefit from adjuvant anthracycline-based chemotherapy, including *ERBB2* (HER2) [22], TIMP1 [23] and *TOP2A* [24]. None of these three markers showed the postulated association with outcome in our cohorts of adjuvant or first-line FAC/FEC treated patients (**Supplementary Table 10**), neither did we observe a correlation between mtDNA content and these markers (**Supplementary Table 11**).

The retrospective nature of this study did not allow us to include patients receiving taxane-based chemotherapy, nowadays often used to treat breast cancer as well. Similar to the methotrexate-based regimen, taxanes induce only low levels of oxidative stress [16] and do not work directly at the DNA level. Thus, we do not expect differences in efficacy in tumours with low or high mtDNA content. Specifically, we reason that tumour cells with low mtDNA content possess a vulnerability to drugs affecting mitochondria. It would be interesting to assess the outcome of patients with low or high tumour mtDNA content when treated with regimen containing chemotherapeutics that induce either mild (such as taxanes) or moderate oxidative stress (such as platins or alkylating agents), or the sequential combination of regimen. But given the common use of taxanes in breast cancer patients and the retrospective nature of our study, validation of our findings as well as exploring the impact of mtDNA content on outcome of taxane-treated patients is warranted.

In conclusion, where it was previously shown that a low mtDNA content in primary breast cancer tumours is associated with a worse prognosis [14], our study here in two well-defined cohorts indicates that breast cancer patients with low mtDNA content in their primary tumour have better outcome from anthracycline-containing chemotherapy, in the adjuvant as well as in the advanced setting. We suggest that the frequently observed decrease in mtDNA content in breast tumours may be exploited by guiding chemotherapeutic regimen decision-making. Larger (prospective) cohorts of uniformly treated patients are necessary to validate our results and to determine the clinical relevance of mtDNA content quantification in cancer.

Materials and methods

Study cohort and sampling

DNA extracts from an earlier retrospective study [25] were selected from our bio-bank at the Erasmus MC (**Supplementary Figure 1**). Tissue specimens from our selection were collected during 1979-1995. The study was approved by the medical ethics committee of the Erasmus MC (MEC 02.953) and conducted in accordance to the Code of Conduct of Federation of Medical Scientific Societies in the Netherlands, and thus provided patient records were anonymized and de-identified prior to analysis. We adhered to the Reporting Recommendations for Tumour Marker Prognostic Studies (REMARK) [26]. Follow-up and tumour response were defined based on the criteria by the International Union Against Cancer (Geneva, Switzerland) [27] and the European Organization for Research and Treatment of Cancer [28] as described previously [25]. Estrogen receptor (ER) status, progesterone receptor (PR) status, and *ERBB2* (HER2) amplification were determined as described before (resp. [29, 30] and [31]). The RT-qPCR Genomic Grade Index (GGI) was determined as described previously [32], which we used as an alternative grading system for histological grading according to Bloom-Richardson since this grading was unavailable for nearly 30% of the samples.

For the adjuvant cohort (REMARK diagram in **Supplementary Figure 1A**), a total of 528 DNA extracts were available from fresh frozen primary breast tumour specimens originating from female patients with lymph node-positive disease (N1 or N2) without distant metastasis at primary diagnosis (M0). None of the patients received neoadjuvant therapy. We selected only samples from patients receiving as adjuvant systemic treatment either cyclophosphamide / methotrexate / 5-fluorouracil (CMF) or 5-fluoracil / anthracycline / cyclophosphamide (FAC or FEC) and did not receive adjuvant hormonal treatment (n = 307) (patient characteristics of complete set in **Supplementary Table 1**). Note that at the time of tumour collection, only patients with node-positive disease but not node-negative

disease received adjuvant systemic chemotherapy in the Netherlands. Next, we selected based on uniformity in the DNA extraction method only the samples extracted with the DNeasy Tissue kit (*Qiagen, Venlo, The Netherlands*, performed as described by supplier) (*n* = 289). To minimize stromal cell contamination, we selected for an invasive tumour cell percentage of at least 50% in the specimen (*n* = 176). One patient ineligible in retrospect was excluded. Thus, mtDNA content was examined in a total of 175 patients. For one patient, a qPCR amplification signal was absent. In the final cohort of 174 patients, 56 received CMF and 118 received FAC/FEC adjuvant systemic therapy. At the end of follow-up, 111 events (distant metastasis) were observed within the group of these 174 patients.

For the advanced cohort (REMARK diagram in **Supplementary Figure 1B**), a total of 231 DNA extracts from fresh frozen primary disease resection specimen were available from female patients who received chemotherapy as first-line treatment for disease recurrence (loco-regional or distant metastasis). We selected only those samples originating from patients receiving either CMF or FAC/FEC as first-line chemotherapy regimen (*n* = 206) (patient characteristics of complete set in **Supplementary Table 2**). We again only selected patients based on uniformity in the DNA extraction method (only the samples extracted with the DNeasy Tissue kit (*Qiagen*), performed as described by supplier) (*n* = 197) and with an invasive tumour cell percentage of at least 50% in the specimen (*n* = 126). Five patients were excluded: no clinical data was available for four patient samples, and one patient who did not receive axillary lymph node dissection was ineligible in retrospect. Thus, mtDNA content was examined in a cohort of 121 patients, of which 44 received CMF and 77 received FAC/FEC first-line chemotherapy. A total of 29 (24%) patients received hormonal consolidation therapy during follow-up and for one patient it was unknown if she had been treated with consolidation therapy. Of the 121 patients, none (0%) received neoadjuvant therapy and 34 (28%) received prior adjuvant chemotherapy (7 CMF, 17 FAC/FEC and 9 cyclophosphamide monotherapy). Overall response was defined as described previously [25], and a total of 69 patients responded to chemotherapy (5 with complete remission, 38 with partial remission, 26 with stable disease > 6 courses) and 51 patients showed no response to chemotherapy (36 with progression, 15 with stable disease ≤ 6 courses) (**Supplementary Table 3**). For one patient the type of response was ambiguous. In our analysis, we right-censored at 18 months follow-up because of low numbers of patients at risk. At the end of follow-up, 104 events (progression) were observed within the group of 121 patients.

mtDNA content

Mitochondrial DNA content was determined as described previously [14]. Briefly, a multiplex real-time quantitative PCR was used which amplifies nuclear *HMBS*

(chr11q23.2-qter) and mitochondrial *MT-TL1* (chrMT 3212–3319). ROX normalized quantification cycle values (Cq [dRn]) were obtained from duplicate runs on the MX3000 or MX3005P qPCR systems (*Agilent Technologies, Waldbronn, Germany*) by the adaptive baseline approach (MxPro v4.10) up to cycle 35 with fixed fluorescence thresholds at 0.004 dRn. We quantified in 1 ng DNA the ratio of mitochondrial DNA opposed to nuclear DNA by the relative quantitation method (2^ΔCq [54]) using the obtained Cq values. Multiplying this ratio by the copy number of *HMBS* (obtained as described below) resulted in the number of mtDNA molecules per cell as mtDNA content. A calibration curve containing a pool of DNA extracts from a different set of fresh frozen tumours was taken along as internal control to monitor the performance of the PCR reactions (**Supplementary Table 4**). Also, no template reactions were taken along in each run and never resulted in an obtained Cq value.

Copy number analysis

Copy number variation of the nuclear encoded *HMBS* gene – which served as a reference to obtain mtDNA content – was obtained using qBiomarker Copy Number assays (*Qiagen*) VPH000-0000000A (multi-copy reference) and VPH111-0594782A (*HMBS* target) as described by the supplier. Briefly, *HMBS* and the multi-copy reference were quantified using SYBR-green based real-time quantitative PCR in 4 ng of DNA. The multi-copy reference is a stable sequence that is minimally affected by local genomic changes, because it appears in the human genome over 40 times. ROX normalized Cq values were obtained as described above but with fixed fluorescence thresholds at 0.2 dRn. In 4 ng of DNA, using the 2^ΔΔCq calibrator genome method, the predicted copy number of *HMBS* was obtained for each individual tumour and corrected for tumour cell percentage. A calibration curve containing a pool of DNA extracts from a different set of fresh frozen tumours was taken along as internal control to monitor the performance of the PCR reactions (**Supplementary Table 4**), of which the dilution containing 4 ng of DNA was used as the calibrator for copy number calculations. Two cell line samples were taken along as high (MDA-MB-468) and low (MDA-MB-134VI) controls for copy number variation, which gave results with respectively median 3.3 (min 2.5 – max 4.5) and median 0.9 (min 0.7 – max 1.0) *HMBS* copies/cell (n = 12 runs). Also, no template reactions were taken along in each run and never resulted in an obtained Cq value.

Statistical analysis

All analyses included the average mtDNA content obtained from the duplicate analysis for each individual sample. Data distribution was tested using Shapiro-Wilk test for normality. Categorical comparisons of grouped clinical variables and the mtDNA content groups were

employed using Fisher's exact test. The association with response rate to chemotherapy was analysed with a logistic regression model to calculate odds ratios (ORs) and 95% confidence intervals (95% CIs). Kaplan-Meier survival plots were used to visualize the differences in time to distant metastasis (adjuvant setting) or time to progression (advanced setting) between mtDNA content groups (dichotomized). In the adjuvant cohort the log-rank test was used to compare survival probability. In the advanced cohort the Peto & Peto modification of the Gehan-Wilcoxon test was used to compare survival probability because most events occurred early on in this cohort. Proportional hazard analyses for distant metastasis-free survival (adjuvant setting) or progression-free survival (advanced setting) were performed using Cox regression methods to calculate hazard ratios (HRs) and their 95% CIs. Univariable analysis was done on the individual variables, for the multivariable analysis we used a base model (including traditional clinicopathological variables [23, 33]) and mtDNA content. In regression analyses, the Wald statistic was used to calculate corresponding P values. Proportionality over time was monitored for each Cox regression using the Schoenfeld residuals with the assumption tested using Chi-squared test, and was never violated ($P > 0.05$). All statistical tests were two-sided, and P values smaller than 0.05 were considered as statistically significant. Analyses were performed using R, version 3.2.3.

Supplementary data

Supplementary data for this article are available online at Clinical Cancer Research (http://clincancerres.aacrjournals.org/).

References

1. Robin, E.D. and R. Wong, *Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells.* J Cell Physiol, 1988. **136**(3): p. 507-513.
2. Wiesner, R.J., J.C. Ruegg, and I. Morano, *Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues.* Biochem Biophys Res Commun, 1992. **183**(2): p. 553-539.
3. Legros, F., et al., *Organization and dynamics of human mitochondrial DNA.* J Cell Sci, 2004. **117**(13): p. 2653-2662.
4. Lee, H.C. and Y.H. Wei, *Mitochondrial biogenesis and mitochondrial DNA maintenance of mammalian cells under oxidative stress.* Int J Biochem Cell Biol, 2005. **37**(4): p. 822-834.
5. Reznik, E., et al., *Mitochondrial DNA copy number variation across human cancers.* Elife, 2016. **5**: e10769.
6. Mambo, E., et al., *Tumor-specific changes in mtDNA content in human cancer.* Int J Cancer, 2005. **116**(6): p. 920-924.
7. Yu, M., et al., *Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients.* IUBMB Life, 2007. **59**(7): p. 450-457.
8. Tseng, L.M., et al., *Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer.* Genes Chromosomes Cancer, 2006. **45**(7): p. 629-638.
9. Fan, A.X., et al., *Mitochondrial DNA content in paired normal and cancerous breast tissue samples from patients with breast cancer.* J Cancer Res Clin Oncol, 2009. **135**(8): p. 983-989.
10. Barekati, Z., et al., *Methylation profile of TP53 regulatory pathway and mtDNA alterations in breast cancer patients lacking TP53 mutations.* Hum Mol Genet, 2010. **19**(15): p. 2936-2946.
11. McMahon, S. and T. LaFramboise, *Mutational patterns in the breast cancer mitochondrial genome, with clinical correlates.* Carcinogenesis, 2014. **35**(5): p. 1046-1054.
12. Bai, R.K., et al., *Mitochondrial DNA content varies with pathological characteristics of breast cancer.* J Oncol, 2011. **2011**: 496189.
13. Hsu, C.W., et al., *Mitochondrial DNA content as a potential marker to predict response to anthracycline in breast cancer patients.* Breast J, 2010. **16**(3): p. 264-270.
14. Weerts, M.J.A., et al., *Mitochondrial DNA content in breast cancer: Impact on in vitro and in vivo phenotype and patient prognosis.* Oncotarget, 2016. **7**: p. 29166-29176.
15. Harris, L.N., et al., *Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline.* J Clin Oncol, 2016. **34**(10): p. 1134-1150.
16. Conklin, K.A., *Chemotherapy-associated oxidative stress: impact on chemotherapeutic effectiveness.* Integr Cancer Ther, 2004. **3**(4): p. 294-300.
17. Ashley, N. and J. Poulton, *Mitochondrial DNA is a direct target of anti-cancer anthracycline drugs.* Biochem Biophys Res Commun, 2009. **378**(3): p. 450-455.
18. Khiati, S., et al., *Mitochondrial topoisomerase I (top1mt) is a novel limiting factor of doxorubicin cardiotoxicity.* Clin Cancer Res, 2014. **20**(18): p. 4873-4881.
19. Mei, H., et al., *Reduced mtDNA copy number increases the sensitivity of tumor cells to chemotherapeutic drugs.* Cell Death Dis, 2015. **6**: e1710.
20. Early Breast Cancer Trialists' Collaborative, G., *Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials.* Lancet, 2005. **365**(9472): p. 1687-1717.
21. Cortopassi, G.A. and N. Arnheim, *Detection of a specific mitochondrial DNA deletion in tissues of older humans.* Nucleic Acids Res, 1990. **18**(23): p. 6927-6933.
22. Gennari, A., et al., *HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials.* J Natl Cancer Inst, 2008. **100**(1): p. 14-20.

23. Schrohl, A.S., et al., *Tumor tissue levels of Tissue Inhibitor of Metalloproteinases-1 (TIMP-1) and outcome following adjuvant chemotherapy in premenopausal lymph node-positive breast cancer patients: A retrospective study.* BMC Cancer, 2009. **9**: 322.
24. Du, Y., et al., *The role of topoisomerase IIalpha in predicting sensitivity to anthracyclines in breast cancer patients: a meta-analysis of published literatures.* Breast Cancer Res Treat, 2011. **129**(3): p. 839-848.
25. Liu, J., et al., *The 29.5 kb APOBEC3B deletion polymorphism is not associated with clinical outcome of breast cancer.* PLoS One, 2016. **11**(8): e0161731.
26. McShane, L.M., et al., *Reporting recommendations for tumor marker prognostic studies (REMARK).* J Natl Cancer Inst, 2005. **97**(16): p. 1180-1184.
27. Hayward, J.L., et al., *Assessment of response to therapy in advanced breast cancer: a project of the Programme on Clinical Oncology of the International Union Against Cancer, Geneva, Switzerland.* Cancer, 1977. **39**(3): p. 1289-1294.
28. European Organization for Research and Treatment of Cancer, B.C.C.G., *Manual for clinical research and treatment in breast cancer (4th edition)*. 2000, Excerpta Medica Almere, the Netherlands. p. 116-117.
29. Foekens, J.A., et al., *Prognostic value of estrogen and progesterone receptors measured by enzyme immunoassays in human breast tumor cytosols.* Cancer Res, 1989. **49**(21): p. 5823-5828.
30. Foekens, J.A., et al., *Relationship of PS2 with response to tamoxifen therapy in patients with recurrent breast cancer.* Br J Cancer, 1994. **70**(6): p. 1217-1223.
31. van Agthoven, T., et al., *Relevance of breast cancer antiestrogen resistance genes in human breast cancer progression and tamoxifen resistance.* J Clin Oncol, 2009. **27**(4): p. 542-549.
32. Toussaint, J., et al., *Improvement of the clinical applicability of the Genomic Grade Index through a qRT-PCR test performed on frozen and formalin-fixed paraffin-embedded tissues.* BMC Genomics, 2009. **10**: 424.
33. Schrohl, A.S., et al., *Primary tumor levels of tissue inhibitor of metalloproteinases-1 are predictive of resistance to chemotherapy in patients with metastatic breast cancer.* Clin Cancer Res, 2006. **12**(23): p. 7054-7058.

CHAPTER 4


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Asp
Ser
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
ND3
Gly
Arg
ND4l
His
Ser
Leu

Mitochondrial RNA expression and variants in association with clinical parameters in primary breast cancers

Marjolein J.A. Weerts | Marcel Smid | John A. Foekens |
Stefan Sleijfer | John W.M. Martens

Cancers 2018, 10(12):500

Abstract

The human mitochondrial DNA (mtDNA) encodes 37 genes, including thirteen proteins essential for the respiratory chain, and two ribosomal RNAs and twenty-two transfer RNAs functioning in the mitochondrial translation apparatus. The total number of mtDNA molecules per cell (mtDNA content) is variable not only between tissue types, but also between tumours and their normal counterparts. For breast cancer specifically, tumours tend to be depleted in their mtDNA content compared to adjacent normal mammary tissue. Additionally, various studies have shown that primary breast tumours harbour somatic variants in their mtDNA. A decrease in mtDNA content or the presence of somatic variants could indicate a reduced mitochondrial function within breast cancer. In this explorative study we aimed to further understand the genomic changes and expression of the mitochondrial genome within breast cancer, by analysing RNA sequencing data of primary breast tumour tissue specimens of 344 cases. We demonstrate that somatic variants detected at the mtRNA level are representative for the somatic variants in the mtDNA. We also show that the number of somatic variants within the mitochondrial transcriptome is not associated with the mutational processes impacting the nuclear genome, but is positively associated with age at diagnosis. Finally, we observe that mitochondrial expression is related to ER status. We conclude that there is a large heterogeneity in somatic mutations of the mitochondrial genome within primary breast tumours, and differences in mitochondrial expression among breast cancer subtypes. The exact impact on metabolic differences and clinical relevance deserves further study.

Introduction

Mitochondria are small organelles involved in multiple cellular processes. They are most renowned for their role in energy production, since they contain their own circular genomic entity encoding proteins essential for the respiratory chain and thereby for generating cellular ATP via oxidative phosphorylation. The human mitochondrial DNA (mtDNA) is gene-dense consisting of ~16,569 base pairs encoding 37 genes: thirteen proteins, and two ribosomal RNAs and twenty-two transfer RNAs functioning in the mitochondrial translation apparatus. Polycistronic transcription of mtDNA is initiated at the non-coding D-loop region, and the resultant precursor transcripts are processed by excision of the transfer RNA genes ("tRNA punctuation model" [1]) generating individual mitochondrial transfer RNA, ribosomal RNA and messenger RNA transcripts. The total number of mtDNA molecules per cell (mtDNA content) is variable between tissue types, and interestingly also between tumours and their normal counterparts [2]. For breast cancer specifically, tumours tend to be depleted in their mtDNA content compared to adjacent normal mammary tissue [2-10], and mtDNA content in breast tumours positively correlates with the expression of mtDNA-encoded genes [11]. Decreased content and expression of mtDNA could indicate a reduced mitochondrial function within breast cancer, in line with the Warburg hypothesis [12] limiting energy production largely to glycolysis. Recently, we have shown mtDNA content to be associated with breast cancer patient outcome [13, 14], underlining the clinical relevance of mitochondria in breast cancer.

Apart from mtDNA content, the significance of somatic mtDNA variants within (breast) cancer is still subject to debate, where the whole spectrum of neutral accumulation, positive selection (advantage) and negative selection (disadvantage) have been postulated. Various studies have shown that primary breast tumours harbour somatic variants in their mtDNA [8, 15, 16], with approximately 70% of the specimens containing at least one single nucleotide variant (SNV, range 1 – 7) and 10% containing at least one small insertion/deletion (INDEL, range 0 – 3). However, these variants do not appear at particular 'hot-spot' positions on the mitochondrial genome, raising doubts about their clinical relevance.

To better understand nucleotide changes in and expression of the mitochondrial genome within primary breast tumours, we investigated here transcriptomic sequencing data within the ICGC consortium [17] and explored how these findings correlate with clinical parameters, providing more insight into the mitochondrial genome as potential biomarker and its clinical relevance in breast cancer.

Results

We evaluated RNA sequencing data of 344 primary breast tumour specimens. After mapping of sequencing reads against the human reference genome, median 15% (IQR 10-23%) of the uniquely mapped reads were assigned to the mitochondrial contig, resulting in median 9,889x read depth (IQR 5,333) of mtDNA.

Somatic variants in mtRNA

Variant calling resulted in a total of 9,063 single nucleotide variants (SNVs) on 1,600 positions and 84 small insertions or deletions (INDELs) on 38 positions of the mitochondrial genome within the 344 cases (**Figure 1**). Since INDELs were only a minority, our

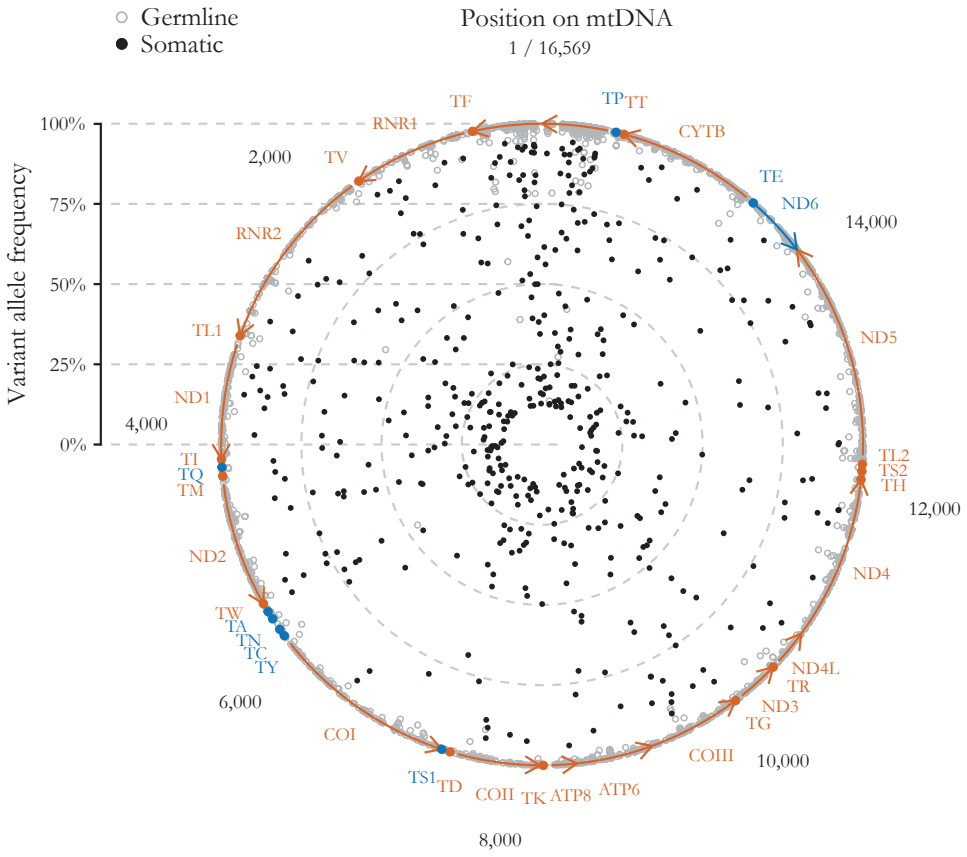


Figure 1 Variants in the mitochondrial RNA of 344 primary breast tumour cases.
Position on the mitochondrial genome (circle) and their variant allele frequency (increasing % from inner-to-outer) of all variants identified in the 344 cases. Somatic or germline origin in respectively closed black or open grey circles. Genes and their direction of transcription (arrows) in red (+ strand) or blue (-strand). Note that variants on position 2617 (known RNA – DNA differences) are not shown.

focus was on the SNVs only. We defined SNVs as somatically acquired tumour variants when not associated with the individual's haplotype (n = 7,235 excluded, 80%) or with heteroplasmic allele frequency of ≤ 95% (n = 917 excluded, 10%). Also, we defined the variants at position 2617 (r.2617a>u and r.2617a>g, present in respectively n = 340 and n = 101 cases) as not tumour-specific because 1) they have been described previously as RNA-DNA differences in blood cells of non-cancer patients [18, 19] and 2) we confirmed their presence in a transcriptomic dataset of normal specimens of various tissue types including breast tissue [20] (**Supplementary Table 4**). After these exclusions, a total of 470 somatic variants on 429 positions were identified.

Our dataset has overlapping cases (n = 165) with the dataset published by Ju et al. [15] concerning somatic mitochondrial variants in tumour and matched normal specimens at the DNA level. This allowed us to directly compare called variants between the two datasets (see also **Supplementary File**) to evaluate presence, classification and allele frequency of variants. Since variants at position 2617 are known RNA-DNA differences (see above) and indeed not called in the DNA dataset, these were not included in this comparison. A total of respectively 3,997 and 4,009 SNVs were called at the RNA and DNA level within the primary tumour specimens of the 165 cases. The majority of the variants were called at both the RNA and DNA level (n = 3,889, respectively 97.3% and 97.0%), whereas a small fraction was only called at either the RNA or the DNA level (respectively n = 108 (2.7%) and n = 120 (3.0%) variants) (**Figure 2**). Of the variants detected at both the RNA and DNA level, only a few (n = 10, 0.3%) had a discrepancy in classification as either 'somatic' or 'germline' (**Figure 2**). Also, good consistency was observed in allele frequency at the RNA and DNA level (linear fit coefficient of 0.92 for all variants and 0.96 for somatic tumour variants). From this we concluded that presence,

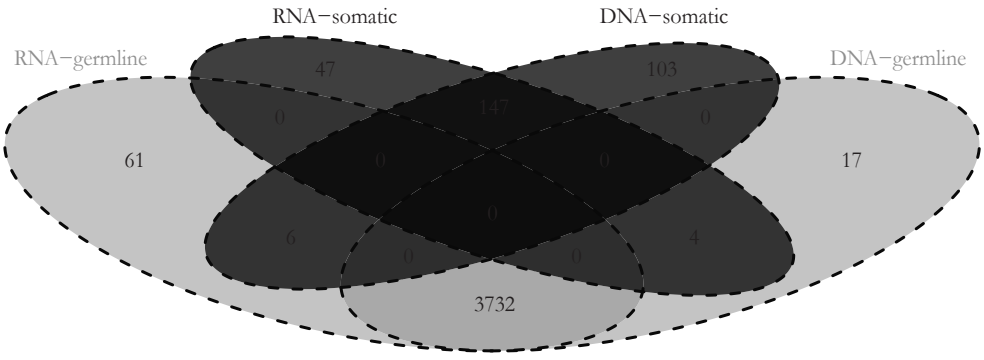


Figure 2 Classification of variants detected in the mitochondrial RNA and in the mitochondrial DNA of 165 primary breast tumour cases.
Venn-diagram depicting classification of variants as either somatic (black) or germline (grey) at the RNA level and the DNA level.

classification and allele frequency of variants was consistent between the RNA and the DNA level (as elaborated on in **Supplementary File**).

We then continued to further decipher the somatic mtRNA variants in our dataset (n = 470 in n = 344 cases). The variant allele frequency of the somatic variants was distributed with a peak at the lower and at the upper end of allele frequencies (**Supplementary Figure S1**). There was no correlation between the variant allele frequency and the percentage of invasive tumour cells in the evaluated specimen (Spearman correlation coefficient rho = 0.03, P = 0.5). The detected somatic variants were distributed along the entire mitochondrial genome (**Figure 1**), with 40 (8.5%) variants located in the tRNA genes, 69 (14.7%) in rRNA genes, 85 (18.1%) in the D-loop, 1 (0.2%) in the non-coding regions, and 275 (58.5%) in the mRNA genes of which 212 (77.1%) had a nonsynonymous effect on the coding amino acid (**Figure 3**). However, relative to their genomic size (9.0% tRNA genes, 15.1% rRNA genes, 6.8% D-loop, 0.4% non-coding and 68.7% mRNA genes) more variants were present in the D-loop and fewer in the mRNA genes (Fisher exact P < 0.001). Also in comparison to the germline variants (variants that were associated with the haplogroup of that individual or with an allele frequency > 95%, n = 8,152) there was a difference in genomic distribution (Fisher's exact P < 0.001) with fewer somatic variants in the D-loop but more in the tRNA and mRNA genes, and an enrichment for somatic nonsynonymous mRNA variants (**Figure 3**). The positions of somatic variants were much more conserved among species compared to the germline variants (Mann-Whitney test P < 0.001), as measured by the fraction of species that harbour the reference

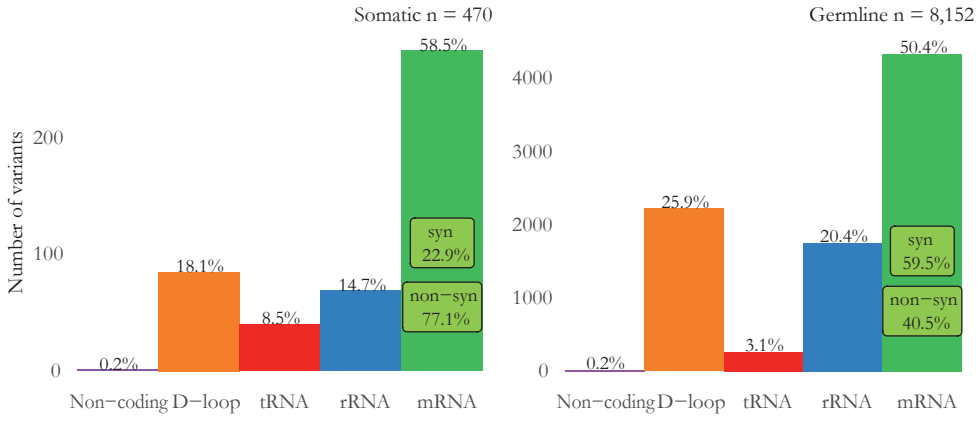


Figure 3 Genomic distribution of mitochondrial RNA variants of 344 primary breast tumour cases.
Genomic distribution is depicted for somatic (left) or germline (right) variants in either non-coding (purple), the D-loop (orange), tRNA (red), rRNA (blue) or mRNA (green) regions of the mitochondrial genome. The percentage of total is indicated at the top of the bars. The percentage of substitutions in the mRNA regions with either a synonymous or non-synonymous effect is indicated within the mRNA bar (light green). Note that variants at position 2617 (known RNA – DNA differences) are not included.

sequence at that position (Conservation Index of respectively median (IQR) 0.93 (0.36) and 0.76 (0.69)). A total of 69 (15%) somatic variants were recurrent and positioned on 28 mitochondrial positions. Majority of the somatic variants (95%) represented the typical replication-coupled mtDNA substitution pattern with predominantly C > T and T > C transitions as described previously [15, 16, 21] in a nucleotide context similar to the germline variants (**Figure 4**). However, compared to the detected germline variants the ratio between C > T and T > C variants is shifted (Fisher exact P < 0.001) with an increased number of C > T transitions among the somatic variants (**Figure 4**).

In the entire cohort, there are 112 (33%) cases with 0 somatic variants, 97 (28%) with 1 somatic variant, and 135 (39%) with more than 1 somatic variant (range 2 to 7). Of the cases with more than 1 somatic variant, 82 (61%) had a difference > 20% allele frequency between variants, indicative for (sub-)clonality.

Somatic mitochondrial variants in relation to somatic variants in the nuclear genome

Next, to gain more insight into the relation between the mutational processes shaping mtDNA and nDNA, we associated the amount of somatic mtRNA variants with the number of somatic variants induced by the known major mutational patterns shaping the nDNA. For this purpose, we obtained for the overlapping cases (n = 268) the number of nDNA variants as published by Nik-Zainal et al. [17]. There was no statistically significant association between the number of somatic mtRNA variants and the total number of somatic variants in the nuclear DNA (Spearman correlation coefficient rho = 0.01, P = 0.8). Next, we combined per case the number of variants in nDNA associated with the mutational processes as described by Nik-Zainal et al. [17]: age-related (signatures 1 and 5), APOBEC-related (signatures 2 and 13) and homologous-recombination deficiency-related (signatures 3 and 8) processes. No statistically significant associations were observed between the number of somatic mtRNA variants and any of these three mutational processes (all Kruskal-Wallis P > 0.2). Note that only two samples within the dataset contained variants associated with mismatch-repair deficiency (signatures 6, 20 and 26), and none of samples contained variants associated with the signatures of unknown aetiology (signatures 17, 18 and 30), as a consequence of which these specific subgroups could not be evaluated.

Mitochondrial gene expression

To estimate the expression and transcript processing of the mitochondrial genome for each case, transcripts per million (TPM, log2-transformed) were calculated for the entire mtDNA and each mitochondrial-encoded gene individually. Expression of the entire mtDNA – normalized against the nuclear genome and thus evaluated as driven by

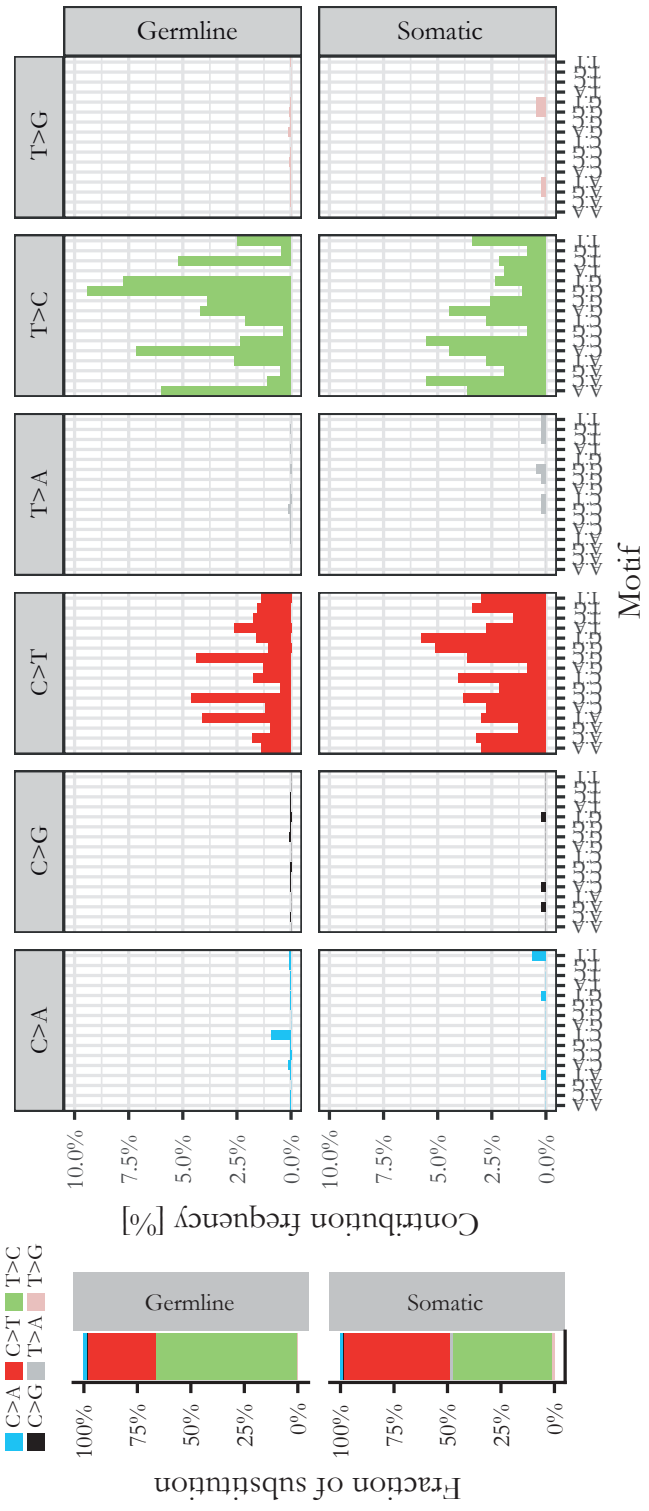


Figure 4 Somatic spectrum of mitochondrial RNA variants of 344 primary breast tumour cases.
The contribution of the six possible base substitutions (C > A in blue, C > G in black, C > T in red, T > A in grey, T > C in green and T > G in pink) (left) and the context of each substitution (bases immediately 5' and 3' to each variant in the reference genome) (right) are depicted for the germline (top) and somatic (bottom) variants (left). Note that variants on position 2617 (known RNA – DNA differences) are not included.

mtDNA content and transcription rate – was high and showed minor variability among the 344 cases (median 19.9210 TPM, IQR 0.0045). Within the 37 mitochondrial-encoded genes – normalized within the mitochondrial genome and thus evaluated as driven by processing of the polycistronic transcripts – the levels for genes encoding tRNAs were lowest (median 12.52 TPM, IQR 1.32), followed by mRNAs (median 15.37 TPM, IQR 0.31) and rRNAs (median 16.83 TPM, IQR 0.48). Most variability

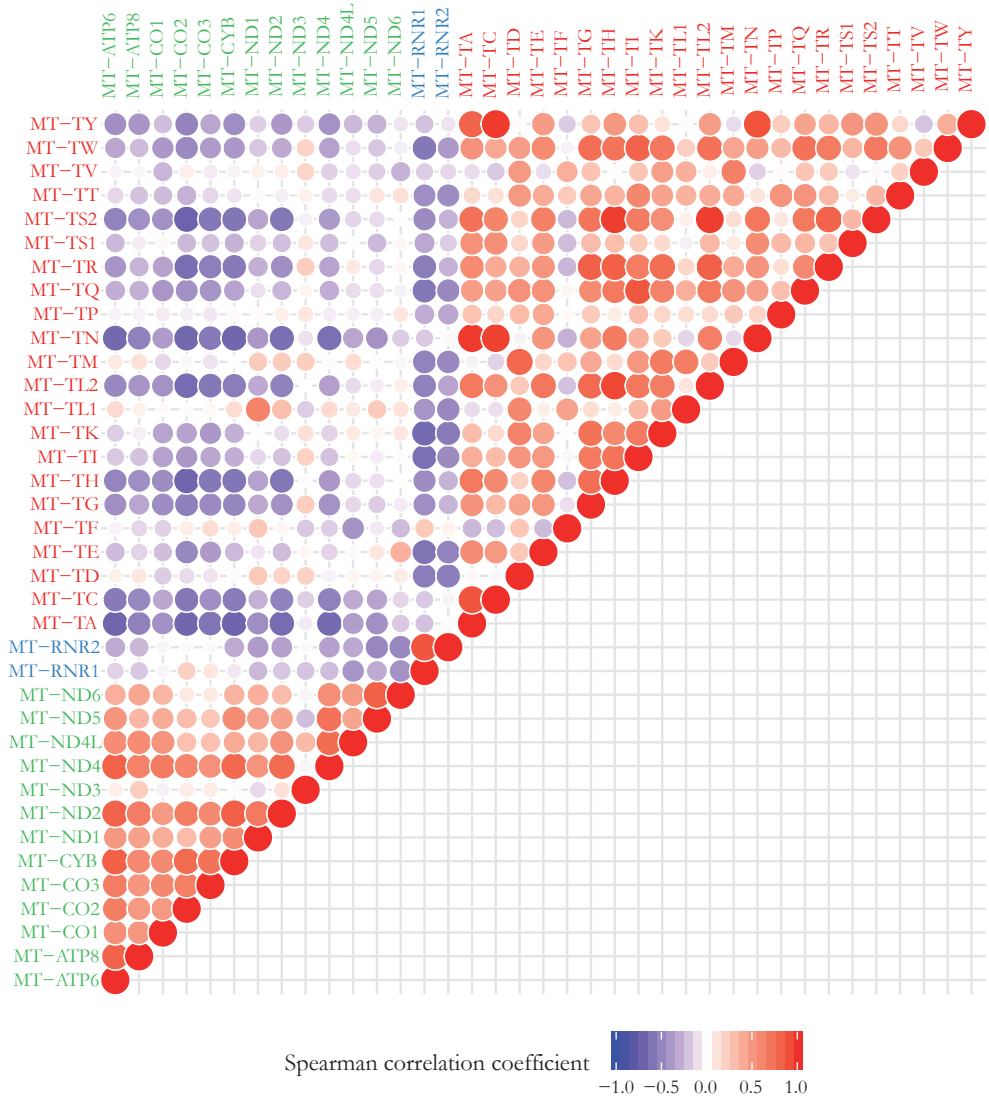


Figure 5 Correlation matrix of expression of all 37 mitochondrial-encoded genes of 344 primary breast tumour cases.

Correlation matrix depicting the Spearman correlation between all 37 mitochondrial-encoded genes (text of tRNA genes in red, rRNA genes in blue, mRNA genes in green). Colour intensity and the size of the circle are proportional to the correlation coefficients.

was observed in levels of tRNAs. Also, distinct correlation clusters were observed between the expression levels of the genes encoding mRNAs, tRNAs and rRNAs, where among genes a positive correlation was present per gene-type, but between different gene-types a negative correlation was present (**Figure 5**). No correlation was observed between the number of mtRNA variants and expression of the entire mtDNA (Spearman correlation coefficient rho = -0.02, P = 0.7).

Association with clinicopathological parameters

Lastly, we explored how these findings correlate with relevant clinical parameters. We analysed the number of somatic mtRNA variants (grouped variable as 0 variants, 1 variant and >1 variant per tumour, **Table 1**) and the expression of the entire mitochondrial contig (continuous variable, **Table 1**) in relation to traditional clinicopathological variables including age at diagnosis (*n* = 291 cases), tumour size (T-stage) (*n* = 216 cases),

Table 1 Association between number of somatic tumour mtRNA variants or expression of the entire mtDNA and clinicopathological variables.

		mtRNA somatic variants				mtRNA expression	
Variable	**No. of cases**	**0 variants**	**1 variant**	**>1 variants**	**P**	**median (IQR) TPM**	**P**
Age					**0.022**[a]		**0.049**[d]
56 (28-85)	291 (100%)	53 (17)	55 (23)	61 (24)		0.11[c]	
unknown	*53*						
Tumour size					0.07[b]		**0.051**[a]
T1 (≤ 2 cm)	76 (35.2%)	33.8%	25.0%	44.4%		19.9202 (0.0043)	
T2 (> 2-5 cm)	109 (50.5%)	47.9%	64.1%	42.0%		19.9207 (0.0045)	
T3 (> 5 cm)	31 (14.4%)	18.3%	10.9%	13.6%		19.9223 (0.0047)	
unknown	*128*						
Grade					0.4[b]		0.1[a]
I	24 (8.5%)	9.9%	12.2%	5.1%		19.9202 (0.0037)	
II	111 (39.4%)	40.7%	35.1%	41.0%		19.9216 (0.0044)	
III	147 (52.1%)	49.5%	52.7%	53.8%		19.9209 (0.0049)	
unknown	*62*						
ER					0.3[b]		**< 0.001**[a]
Negative	81 (27.8%)	21.7%	31.2%	30.3%		19.9196 (0.0050)	
Positive	210 (72.2%)	78.3%	68.8%	69.7%		19.9216 (0.0041)	
unknown	*53*						
PR					0.5[b]		**0.006**[a]
Negative	102 (35.4%)	31.5%	40.5%	35.0%		19.9204 (0.0048)	
Positive	186 (64.6%)	68.5%	59.5%	65.0%		19.9215 (0.0042)	
unknown	*56*						

For each subgroup within the clinicopathological variable, the number of cases and either the fraction of patients within the mtRNA somatic variant groups (0, 1 or more than 1) or the mtRNA expression (TPM, log2 transformed) is indicated. [a] Kruskal-Wallis (multiple groups) or Mann-Whitney (two groups) P value. [b] Fisher exact P value. [c] Spearman correlation coefficient. [d] Spearman correlation P value.

pathological grade (n = 282 cases), estrogen receptor (ER) status (n = 291 cases) and progesterone receptor (PR) status (n = 288 cases). Due to the low numbers of patients with HER2-amplified (n = 2 cases) and presenting with metastases at primary diagnosis (n = 3 cases), these clinicopathological variables were not evaluated. Age at diagnosis was statistically significant associated with both the number of somatic mtRNA variants (Kruskal-Wallis P = 0.022) and expression of the entire mtDNA (Spearman correlation coefficient rho = 0.11, P = 0.049), where a higher age corresponded to more somatic mtRNA variants and higher expression of the entire mtDNA. Also, a highly statistically significant association was observed between expression of the entire mtDNA and hormone receptor status (as evaluated at the protein level), with increased mtDNA expression in the ER-positive and in the PR-positive tumours (respectively Mann-Whitney U test P < 0.001 and P = 0.006). In fact, also a significant correlation was observed between expression of the entire mtDNA and RNA expression of *ESR1* or *PGR* (respectively Spearman correlation coefficient rho = 0.19 P < 0.001 and rho = 0.17 P = 0.001, n = 344 and n = 342 cases).

Discussion

In this work, we explored genomic changes in and expression of the mitochondrial genome within primary breast tumours, and their correlation with clinicopathological variables.

Within our breast tumour dataset, the fraction of reads mapping to the mitochondrial contig of the reference genome (median 15%) is in line with previous findings in non-tumorous breast samples: within the Illumina Body Tissue Atlas ~15% of the sequencing reads mapped to the mitochondrial genome (n = 1) [22], and within the Genotype-Tissue Expression (GTEx) Consortium ~15-20% of the transcriptional output was of mitochondrial origin (n = 27) [23]. This indicates that although the expression of the mitochondrial genome has been shown to be decreased in breast tumours compared to tumour-adjacent normal mammary tissue [11], the extent to which this occurs is less extreme than observed among tissue types (e.g. a much lower fraction of mitochondrial reads in blood (< 5%) or much higher fraction in kidney (>50%) [23]). Nevertheless, we observed an association between expression of the entire mtDNA and ER status (measured at protein-level), with marginally higher expression in ER-positive tumours and a similar observation for PR status (protein-level) (**Table 1**). In addition, also RNA expression of *ESR1* and *PGR* was positively correlated with expression of the entire mitochondrial contig. The relation between expression of mtDNA and clinicopathological parameters has not been evaluated by others, but when we associated the data reported by Reznik et al [11] on mtRNA expression within the TCGA-BRCA dataset (n = 656 cases) we observe a

similar correlation for ER status (Kruskal-Wallis P = 0.006, **Supplementary Table 5**) and none for the other clinicopathological variables (all P > 0.05 **Supplementary Table 5**). In pre-clinical models, there appears to be a link between ER and mitochondrial activity: exposure to estrogens increases mitochondrial expression and oxygen consumption in ER-positive [24, 25] but not in ER-negative breast cancer cells [25]. Similarly, ER-negative breast cancer cell lines show lower mitochondrial respiration and a stronger dependency on glycolysis in comparison to ER-positive breast cancer cells [26]. Unfortunately, measurements on mitochondrial activity comparing ER-positive and ER-negative clinical specimens are to our knowledge not reported in the literature, and thus the effect of differences in *ESR1* levels on mitochondrial activity in primary breast tumours remains currently unknown. Interestingly, uptake values of fluorodeoxyglucose (FDG) in positron emission tomography (PET) – a visualization of glucose uptake reflecting the increased rate of glycolysis in the tumour – appears to be higher in ER-negative cases [27-33], indicative that indeed metabolic differences are present between the subtypes. Additional studies should be performed to identify if there are differences in mitochondrial function among breast cancer subtypes and the potential clinical relevance of these findings, such as predictive and prognostic potential.

We also observed distinctive clustering of tRNA genes, which is in line with the tRNA punctuation model: when processing the polycistronic transcripts, tRNA genes are excised and due their small size (< 75 base pairs) tRNAs are more likely to be lost during the RNA extraction and/or library preparation procedures, whereas the mRNA and rRNA genes are retained (> 200 base pairs). Notably, we did not observe differences in this distinct pattern between the ER-positive and the ER-negative cases (**Supplementary Figures S2 and S3**), and thus the processing of the polycistronic transcripts does not seem to differ between these two subtypes.

Our findings on the number, genomic distribution, and substitution pattern of mtDNA variants within the mitochondrial transcriptome are in line with previous studies on variants within the mitochondrial genome in other cancer types [8, 15, 16, 21, 34, 35] (see also **Supplementary File**). We observe an increased number of somatic variants in the D-loop and fewer in mRNA genes than expected by genomic size, which might be explained by the gene-dense constitution of mtDNA: variants in the D-loop potentially have less destructive effects whereas variants in the mRNA genes might have detrimental effects on the function of the oxidative phosphorylation system, and thus will be selected against. However, compared to germline variants in our dataset there are fewer variants in the D-loop and more in the tRNA and mRNA genes, and enrichment for nonsynonymous variants. This might be explained by the typical mutation pattern shaping mtDNA, which has been shaping the germline variants and thus the trivial positions have already been altered, as suggested by Ju et al [15]. In line with this, the

conservation of variants among species – the fraction of species that harbour the reference sequence at that position – was much higher for somatic variants than for the germline variants, which can be explained by the same hypothesis. Adding to this, compared to the detected germline variants there is an increased number of C > T transitions among the somatic variants (**Figure 4**).

Adjusting variant allele frequency to account for sample purity (percentage of tumour cells within the specimen) is often applied for nuclear-encoded genes to obtain information on the allele frequency of variants in the tumour cells. However, this is not possible for mtDNA variants in tumour tissue specimens: the number of mtDNA molecules per cell largely varies among cell types and thus the non-tumour cells present in the specimen do not have the regular two copies as the nuclear genome would have, but contain multiple mtDNA copies of an unknown number. As a result, whereas allele frequency of variants could give information on possible constraints on variants, we did not perform analysis on it since it is impossible to estimate the actual allele frequency of variants in the mitochondria of tumour cells. Nevertheless, we show that majority of the samples with more than 1 somatic variant harbour a difference in variant allele frequency between variants, indicative for (sub-)clonality. This corresponds to the hypothesis that mtDNA variants are either expanded or lost [36] and that the mutations occur separated in time [15].

Also noteworthy is that with the current methodologies applied by us and by others – namely the use of non-micro dissected tumour specimens and blood as matched normal DNA – we cannot be completely sure that the detected somatic mtDNA mutations are tumour-specific. First, tumour tissue specimens consist of multiple cell types, including the tumour cells but also non-neoplastic cells such as immune cells and cells from the mammary epithelium, all with variable mtDNA content. Secondly, (somatic) mtDNA variant heteroplasmy patterns can differ within an individual across tissues [37-40]. Thus, the somatic variants were either acquired in the tumour, the normal somatic epithelium, or even in other cell types present within the specimen.

We did not observe associations between the number of somatic mtRNA variants and the three major mutational processes shaping the nDNA within breast tumours. This is in line with the hypothesis that mutations within the mitochondrial genome are mainly due to fidelity of the mitochondrial polymerase [41] and thereby hardly due to exogenous factors [15]. Accordingly, in our evaluation of associations with clinicopathological parameters we observed a statistically significant association between the number of mtRNA somatic variants and age at diagnosis. Previous work on somatic variants at the DNA level also revealed a correlation with older age of diagnosis (n = 381 [15] and n = 58 cases [34]). Previous work in a small cohort also showed associations between number of somatic variants in mtDNA and higher TNM and higher histological grade

(*n* = 58 cases [34]), which we did not observe. Please note that there are differences in the composition of the cohorts; our dataset does not exactly represent the breast cancer population as seen in daily practice, with an underrepresentation of *ERBB2*-amplified cases (**Supplementary Table 1**).

By using data at the RNA level, we intended to minimize the interference of NUMTs with evaluation of mtDNA expression and variant calling, since their expression in the nucleus is negligibly low [11, 42]. Especially in defining heteroplasmic mtDNA variants in DNA data, NUMTs have been shown to be a complicating issue with non-identical positions misinterpreted as heteroplasmic variants [43-47]. Note that we do observe a few heteroplasmic variants at the DNA-only level (**Supplementary File**). However, using data at the RNA level comes with the trade-off that only variants in expressed regions are detected and thus variants in non-expressed regions are missed. Since mtDNA is a gene-dense entity, we estimate that the number of missed variants should be low. Indeed, in our direct comparison of samples with variants at the RNA and DNA level, we show that this is maximally ~3% of the variants (DNA-only variants). Similar to these findings, the comparison by Stewart et al [16] on somatic variants at the RNA and DNA level showed 7 of the 130 variants (5%) detected at only the DNA level within their set of 100 breast cancer specimens. Another trade-off using RNA is the additional step to generate cDNA, which might induce false positive calls by mistakes of the reverse transcriptase. Again based on our direct comparison of samples with variants at the RNA and DNA level, the number of false positives is maximally 3% of the detected variants (RNA-only variants). Though, besides false positives, these RNA-only variants might actually be RNA-DNA differences for example caused by RNA-editing [48], or true variants not called at the DNA level.

To conclude, in this explorative study on the role of mtRNA in breast cancer, we found that somatic variants at the DNA level are reflected at the RNA level with no hotspot mutations and great heterogeneity across tumours. We confirm that the number of somatic variants within the mitochondrial transcriptome is not associated with the mutational processes shaping the nuclear genome but instead, is associated with age of diagnosis. Furthermore, we show that mitochondrial expression is related to ER status. The exact consequence of the observed differences in mtRNA expression and the detected somatic variants on metabolism and clinical outcome warrants further study.

Materials and methods

Data

We studied all patients with RNA sequencing data within the ICGC BASIS consortium, of which the cohort has been described previously [17] and data deposited in the European-

Genome Phenome Archive (accession code EGAS00001001178). Briefly, for a total of 348 primary breast tumours we generated duplex-specific nuclease-based RNA sequencing data. Four samples were excluded from analyses due to potential cross-contamination (see below). We did not apply a threshold on tumour cell percentage within the specimen for inclusion in this study. Clinicopathological data and the nuclear somatic mutation catalogue were obtained from the Supplementary Tables as provided by Nik-Zainal et al [17]. Expression levels of *ESR1*, *PGR* (quantile normalized FPKM, log2 transformed) were obtained as described previously [49]. A complete dataset on all variables used in our analyses is provided in **Supplementary Table 1**. In addition, we used publically available RNA sequencing data of twelve human tissue specimens obtained via a similar sequencing approach [20], that has been deposited in NCBI's Gene Expression Omnibus (GEO) (accession code GSE45326). Also, we used the mtDNA variants called by Ju et al [15] from whole-genome or whole-exome sequencing data of DNA from the primary breast tumour specimens and matched normal tissue specimens as provided in their Supplementary Tables.

Bioinformatics

Sequencing reads were aligned using STAR v2.4.2.a [50] against the Genome Reference Consortium Human Build 38 (GRCh38, GenBank assembly GCA_000001405.15), which contains as the mitochondrial contig the revised Cambridge Reference Sequence (rCRS). Only non-duplicated uniquely mapped reads on mtDNA were used for further analysis, to avoid the potential use of improper assigned nuclear insertions of mitochondrial origin (NUMTs, mitochondrial pseudogenes). Note that RNA expression of NUMTs has been shown to be absent or negligibly low [11, 42]. Total read depth was estimated based on the read length (75 nucleotides) and mtDNA size (16,569 nucleotides). FeatureCounts v 1.4.6 [51] was used to count mapped reads using mtDNA as the meta-feature and each genomic region (13 mRNAs, 22 tRNAs, 2 rRNAs) as the features, allowing multi-overlapping reads (-O) because of the polycistronic nature of mitochondrial RNA transcripts. We normalized read counts to transcripts per million (TPM) for the entire mitochondrial contig (mtDNA read counts versus total read counts assigned to genes in GRCh38, defined as entire mtDNA levels) and for each mitochondrial-encoded gene (gene read counts versus total mtDNA read counts, defined as <*gene*> levels). In this way, the TPM for the entire mtDNA represents the total amount of mtRNA influenced by both mtDNA content, transcription rate and transcript stability, whereas the TPM for each mitochondrial-encoded gene represents the variation in gene expression driven by processing of the polycistronic transcripts and transcript stability [52]. A complete dataset of all expression levels is provided in **Supplementary Table 2**. Variants alternative

to rCRS were called using GATK HaplotypeCaller 3.4-46-gbc02625 [53] using default settings (including downsampling_type = BY_SAMPLE, downsample_to_coverage = 500, standard_min_confidence_threshold_for_calling = 20). In this way, maximum depth of coverage is controlled at each locus, resulting in a more even coverage of variants between the samples. Hard-filtering was applied to the called variants for quality by depth (QD > 2), alternative depth (AD of ALT > 10) and strand odds ratio (variants with allele frequency ≤ 95% i.e. heteroplasmic variants: SOD < 4 for SNVs and SOD < 10 for INDELs; variants with allele frequency > 95% i.e. (near) homoplasmic: no filtering). In this way, the allele frequency of detected variants was high and confident enough to be a true variant and likely no sequencing errors or PCR mistakes. In addition, after visual inspection of variants (Integrative Genomics Viewer [54, 55]), potential false positive calls in challenging regions were excluded: positions surrounding the homopolymer region 301-315 ("D310"), positions 512-513 due to a repetitive sequence, alternative C calls at positions 16182-16183 and 16189 due to polyC sequences, and alternative A at positions 4264, 5513 and 12138-12139 due to polyA sequences. A complete dataset of all remaining variants is provided in **Supplementary Table 3**. All remaining single nucleotide variants were used in a nucleotide BLAST against the human reference sequence (NCBI's nucleotide web blast, https://blast.ncbi.nlm.nih.gov) with the surrounding reference sequence (30 bases 5' and 30 bases 3') to uncover potential NUMT events, but none were recovered. The conservation index (45 species conservation) for the protein-coding genes, tRNAs and rRNAs were obtained via SNV Query in Mitomaster [56]. The haplotype of each case was estimated by using the heteroplasmic and homoplasmic variants in HaploGrep v2 [57]. Sample cross-contamination was estimated using only the heteroplasmic variants (allele frequency ≤ 95%) in haplotype assignment. This identified four samples with heteroplasmic contamination of another haplotype, therefore these samples were excluded from analyses. Sample mismatch between cases with variants called in both RNA (our dataset) and DNA (dataset Ju et al [15]) sequencing data (n = 168) was estimated by haplotyping based on all near-homoplasmic variants (allele frequency > 95%), and comparison of the obtained haplogroup. Mismatch was observed for 13 patients, but after manual inspection specificity could be confirmed for 10 patients by the presence of private variants. Two patients with a clear mismatch, and one patient ambiguous in mismatch, were excluded from the RNA-DNA comparison analyses (n = 165 remaining).

Statistics

Performed statistical tests are reported in the results section. All statistical tests were two-sided, and P values smaller than 0.05 were considered as statistically significant.

Outliers data points in boxplots are defined as Q1-1.5*IQR or Q3+1.5*IQR. Analyses were performed in R version 3.3.2 (https://cran.r-project.org). Data analyses included usage of the following packages: the set of tidyverse, ggcorplot, SomaticSignatures [58] and VennDiagram [59].

Supplementary data

Supplementary data for this article are available online at Cancers (https://www.mdpi.com/journal/cancers).

References

1. Ojala, D., J. Montoya, and G. Attardi, *tRNA punctuation model of RNA processing in human mitochondria.* Nature, 1981. **290**(5806): p. 470-474.
2. Reznik, E., et al., *Mitochondrial DNA copy number variation across human cancers.* Elife, 2016. **5**: e10769.
3. Mambo, E., et al., *Tumor-specific changes in mtDNA content in human cancer.* Int J Cancer, 2005. **116**(6): p. 920-924.
4. Yu, M., et al., *Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients.* IUBMB Life, 2007. **59**(7): p. 450-457.
5. Tseng, L.M., et al., *Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer.* Genes Chromosomes Cancer, 2006. **45**(7): p. 629-638.
6. Fan, A.X., et al., *Mitochondrial DNA content in paired normal and cancerous breast tissue samples from patients with breast cancer.* J Cancer Res Clin Oncol, 2009. **135**(8): p. 983-989.
7. Barekati, Z., et al., *Methylation profile of TP53 regulatory pathway and mtDNA alterations in breast cancer patients lacking TP53 mutations.* Hum Mol Genet, 2010. **19**(15): p. 2936-2946.
8. McMahon, S. and T. LaFramboise, *Mutational patterns in the breast cancer mitochondrial genome, with clinical correlates.* Carcinogenesis, 2014. **35**(5): p. 1046-1054.
9. Bai, R.K., et al., *Mitochondrial DNA content varies with pathological characteristics of breast cancer.* J Oncol, 2011. **2011**: 496189.
10. Hsu, C.W., et al., *Mitochondrial DNA content as a potential marker to predict response to anthracycline in breast cancer patients.* Breast J, 2010. **16**(3): p. 264-270.
11. Reznik, E., et al., *Mitochondrial respiratory gene expression is suppressed in many cancers.* Elife, 2017. **6**: e21592.
12. Warburg, O., *On the origin of cancer cells.* Science, 1956. **123**(3191): p. 309-314.
13. Weerts, M.J.A., et al., *Mitochondrial DNA content in breast cancer: Impact on in vitro and in vivo phenotype and patient prognosis.* Oncotarget, 2016. **7**: p. 29166-29176.
14. Weerts, M.J.A., et al., *Low tumor mitochondrial DNA content is associated with better outcome in breast cancer patients receiving anthracycline-based chemotherapy.* Clin Cancer Res, 2017. **23**(16): p. 4735-4743.
15. Ju, Y.S., et al., *Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer.* Elife, 2014. **3**: e02935.
16. Stewart, J.B., et al., *Simultaneous DNA and RNA mapping of somatic mitochondrial mutations across diverse human cancers.* PLoS Genet, 2015. **11**(6): e1005333.
17. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.
18. Bar-Yaacov, D., et al., *RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA.* Genome Res, 2013. **23**(11): p. 1789-1796.
19. Hodgkinson, A., et al., *High-resolution genomic analysis of human mitochondrial RNA sequence variation.* Science, 2014. **344**(6182): p. 413-415.
20. Nielsen, M.M., et al., *Identification of expressed and conserved human noncoding RNAs.* RNA, 2014. **20**(2): p. 236-251.
21. Grandhi, S., et al., *Heteroplasmic shifts in tumor mitochondrial genomes reveal tissue-specific signals of relaxed and positive selection.* Hum Mol Genet, 2017. **26**(15): p. 2912-2922.
22. Mercer, T.R., et al., *The human mitochondrial transcriptome.* Cell, 2011. **146**(4): p. 645-658.
23. Mele, M., et al., *Human genomics. The human transcriptome across tissues and individuals.* Science, 2015. **348**(6235): p. 660-665.
24. Chen, J.Q., et al., *Mitochondrial localization of ERalpha and ERbeta in human MCF7 cells.* Am J Physiol Endocrinol Metab, 2004. **286**(6): E1011-22.
25. Mattingly, K.A., et al., *Estradiol stimulates transcription of nuclear respiratory factor-1 and increases mitochondrial biogenesis.* Mol Endocrinol, 2008. **22**(3): p. 609-622.

26. Pelicano, H., et al., *Mitochondrial dysfunction in some triple-negative breast cancer cell lines: role of mTOR pathway and therapeutic potential.* Breast Cancer Res, 2014. **16**(5): 434.
27. Wang, C.L., et al., *Positron emission mammography: correlation of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 status and 18F-FDG.* AJR Am J Roentgenol, 2011. **197**(2): W247-55.
28. Yoon, H.J., et al., *Correlation of breast cancer subtypes, based on estrogen receptor, progesterone receptor, and HER2, with functional imaging parameters from (6)(8)Ga-RGD PET/CT and (1)(8) F-FDG PET/CT.* Eur J Nucl Med Mol Imaging, 2014. **41**(8): p. 1534-1543.
29. Gil-Rendo, A., et al., *Association between [18F]fluorodeoxyglucose uptake and prognostic parameters in breast cancer.* Br J Surg, 2009. **96**(2): p. 166-170.
30. Ikenaga, N., et al., *Standardized uptake values for breast carcinomas assessed by fluorodeoxyglucose-positron emission tomography correlate with prognostic factors.* Am Surg, 2007. **73**(11): p. 1151-1157.
31. Mavi, A., et al., *The effects of estrogen, progesterone, and C-erbB-2 receptor states on 18F-FDG uptake of primary breast cancer lesions.* J Nucl Med, 2007. **48**(8): p. 1266-1272.
32. Nakajo, M., et al., *FDG PET/CT and diffusion-weighted imaging for breast cancer: prognostic value of maximum standardized uptake values and apparent diffusion coefficient values of the primary lesion.* Eur J Nucl Med Mol Imaging, 2010. **37**(11): p. 2011-2020.
33. Osborne, J.R., et al., *18F-FDG PET of locally invasive breast cancer and association of estrogen receptor status with standardized uptake value: microarray and immunohistochemical analysis.* J Nucl Med, 2010. **51**(4): p. 543-550.
34. Tseng, L.M., et al., *Somatic mutations of the mitochondrial genome in human breast cancers.* Genes Chromosomes Cancer, 2011. **50**(10): p. 800-811.
35. Tan, D.J., R.K. Bai, and L.J. Wong, *Comprehensive scanning of somatic mitochondrial DNA mutations in breast cancer.* Cancer Res, 2002. **62**(4): p. 972-976.
36. Coller, H.A., et al., *High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection.* Nat Genet, 2001. **28**(2): p. 147-150.
37. Samuels, D.C., et al., *Recurrent tissue-specific mtDNA mutations are common in humans.* PLoS Genet, 2013. **9**(11): e1003929.
38. He, Y., et al., *Heteroplasmic mitochondrial DNA mutations in normal and tumour cells.* Nature, 2010. **464**(7288): p. 610-614.
39. Li, M.K., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(8): p. 2491-2496.
40. Calloway, C.D., et al., *The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age.* Am J Hum Genet, 2000. **66**(4): p. 1384-1397.
41. Johnson, A.A. and K.A. Johnson, *Exonuclease proofreading by human mitochondrial DNA polymerase.* J Biol Chem, 2001. **276**(41): p. 38097-38107.
42. Collura, R.V., M.R. Auerbach, and C.B. Stewart, *A quick, direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes.* Curr Biol, 1996. **6**(10): p. 1337-1339.
43. Ramos, A., et al., *Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies.* Mitochondrion, 2011. **11**(6): p. 946-953.
44. Parr, R.L., et al., *The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation.* BMC Genomics, 2006. **7**: 185.
45. Parfait, B., et al., *Co-amplification of nuclear pseudogenes and assessment of heteroplasmy of mitochondrial DNA mutations.* Biochem Biophys Res Commun, 1998. **247**(1): p. 57-59.
46. Albayrak, L., et al., *The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome.* BMC Genomics, 2016. **17**(1): 1017.
47. Hazkani-Covo, E., R.M. Zeller, and W. Martin, *Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes.* PLoS Genet, 2010. **6**(2): e1000834.

48. Knoop, V., *When you can't trust the DNA: RNA editing changes transcript sequences.* Cell Mol Life Sci, 2011. **68**(4): p. 567-586.
49. Smid, M., et al., *Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration.* Nat Commun, 2016. **7**: 12910.
50. Dobin, A. and T.R. Gingeras, *Mapping RNA-seq reads with STAR.* Curr Protoc Bioinformatics, 2015. **51**: 11.14.1-19.
51. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-930.
52. Idaghdour, Y. and A. Hodgkinson, *Integrated genomic analysis of mitochondrial RNA processing in human cancers.* Genome Med, 2017. **9**(1): 36.
53. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: 11.10.1-33.
54. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-192.
55. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-26.
56. Lott, M.T., et al., *mtDNA variation and analysis using Mitomap and Mitomaster.* Curr Protoc Bioinformatics, 2013. **44**: 1.23.1-26.
57. Weissensteiner, H., et al., *HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing.* Nucleic Acids Res, 2016. **44**(W1): W58-63.
58. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants.* Bioinformatics, 2015. **31**(22): p. 3673-3675.
59. Chen, H. and P.C. Boutros, *VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.* BMC Bioinformatics, 2011. **12**: 35.

CHAPTER 5


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Ser
Asp
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Ser
Leu

Somatic tumour mutations detected by targeted next generation sequencing in minute amounts of serum-derived cell-free DNA

Marjolein J.A. Weerts* | Ronald van Marion* | Jean C.A. Helmijr | Corine M. Beaufort | Niels M.G. Krol | Anita M.A.C. Trapman-Jansen | Winand N.M. Dinjens | Stefan Sleijfer | Maurice P.H.M. Jansen | John W.M. Martens

* These authors contributed equally to this manuscript

Scientific Reports 2017; 7:2136

Abstract

The use of blood-circulating cell-free DNA (cfDNA) as 'liquid-biopsy' is explored worldwide, with hopes for its potential in providing prognostic or predictive information in cancer treatment. In exploring cfDNA, valuable repositories are biobanks containing material collected over time, however these retrospective cohorts have restrictive resources. In this study, we aimed to detect tumour-specific mutations in only minute amounts of serum-derived cfDNA by using a targeted next generation sequencing (NGS) approach. In a retrospective cohort of ten metastatic breast cancer patients, we profiled DNA from primary tumour tissue (fresh frozen), tumour-adjacent normal tissue (formalin-fixed paraffin embedded), and three consecutive serum samples (frozen). Our presented workflow includes comparisons with matched normal DNA or *in silico* reference DNA to discriminate germline from somatic variants, validation of variants through the detection in at least two DNA samples of an individual, and the use of public databases on variants. By our workflow, we were able to detect a total of four variants traceable as circulating tumour DNA (ctDNA) in the sera of three of the ten patients.

Introduction

Blood-circulating nucleic acids are extracellular (cell-free, cf) residing DNA or RNA molecules (cfDNA or cfRNA) that likely originate from apoptotic or necrotic cells, or are actively released [1]. In cancer patients, cfDNA may harbour somatically derived tumour-specific mutations reflecting the genomic characteristics of an individual's cancer, such as single nucleotide variants or structural rearrangements [2, 3]. The use of cfDNA as a so-called 'liquid-biopsy' is being explored worldwide, with hopes for its potential in providing prognostic or predictive information. In exploring the potential of cfDNA in oncology, valuable repositories are biobanks containing material collected over time. However, these retrospective cohorts have restrictive resources, including suboptimal specimen preservations such as formalin fixed and paraffin embedded (FFPE) specimens or limiting amounts of DNA available for profiling, challenging the analysis of these cohorts.

In this study, we aimed to detect tumour-specific mutations in only minute amounts of serum-derived cfDNA. We present a NGS workflow using a custom 45 gene sequencing panel on the Ion PGM system applied to a retrospective cohort of ten metastatic breast cancer patients. This workflow includes comparisons with FFPE matched normal DNA from tumour-adjacent histologically normal mammary epithelium or *in silico* reference DNA to discriminate germline from somatic variants. To evaluate the detection of variants in minute quantities of DNA, we initially compared, for two patients, the performance of our method on standard amounts of primary tumour DNA with minute counterparts. Finally, for all ten patients, we identified variants in their primary tumour DNA (standard DNA quantities) and three consecutive serum samples (minute DNA quantities)

Results

Performance of the custom amplification-based targeted NGS panel

Using our custom panel we sequenced a total of forty-six DNA samples, derived from fresh frozen (FF) primary breast carcinoma specimens (tumour), FF serum (cfDNA) and FFPE tumour-adjacent normal mammary epithelial specimens (matched normal). Primary tumour (n = 10) and matched normal (n = 6) DNA samples were analysed using standard amounts of DNA input. The cfDNA samples (n = 30) were analysed using minute amounts of DNA input (median 387 pg, interquartile range IQR 265-445 pg) as well as two replicates of primary tumour DNA for two patients (n = 4) (250 pg).

For performance assessment, we analysed the following parameters of the data. Before mapping we analysed 1) read length distribution and 2) per sequence GC content of the generated reads. After mapping against the reference genome we analysed 3) read for each

amplicon, 4) percentage of reads mapped to the targeted regions relative to all mapped reads and 5) read depth of those mapped reads. Finally, after calling single nucleotide variants deviant from the reference genome we analysed 6) the percentage of bases called at the required read depth.

First, the distribution in read length of the reads of each sample was compared to the expected distribution in read length based on amplicon size of the panel (**Supplementary Figure S1A**). The highest peak of read length was at the expected 120 base pairs (bp) for the primary tumour DNA sequenced at standard and at minute amounts as well as for the cfDNA sequenced at minute amounts, whereas the matched normal DNA had its highest peak at 90 bp (Kruskal Wallis $P < 0.001$). Second, we analysed the GC content of the reads for each sample (**Supplementary Figure S1B**). The highest peak for GC content was at the expected 50% for the primary tumour DNA sequenced at standard and minute amounts and the cfDNA sequenced at minute amounts, whereas the matched normal DNA had the highest peak at 40% (Kruskal Wallis $P = 0.002$). Third, we analysed the amplicon performance for each sample type by the median reads per amplicon. The required read depth of at least 20x was obtained for 3019/3106 (97.2%) amplicons in matched normal DNA. The required read depth of at least 100x was obtained for 2973/3106 (95.7%) in the primary tumour DNA at standard amounts, 2953/3106 (95.1%) at minute amounts, and 2947/3106 (94.9%) for cfDNA at minute amounts. A total of 66/3106 amplicons (2.1%) did not reach the required read depths in all four sample types covering regions in 25 genes. Fourth, we determined the percentage of reads mapped to the targeted regions relative to all mapped reads. This percentage was for matched normal DNA sequenced at standard amounts (90.4%, IQR 76.2-93.1%) approximately 5% less compared to primary tumour DNA sequenced at standard amounts (95.7%, IQR 95.6-95.8%) or cfDNA sequenced at minute amounts (94.5%, IQR 94.3-94.7%) (Mann Whitney both $P < 0.001$). This percentage was for cfDNA approximately 1% less than primary tumour DNA sequenced at standard amounts (Mann Whitney $P < 0.001$). The percentage of mapped reads at the targeted regions did not differ between primary tumour DNA sequenced at minute (95.2%, IQR 95.2-95.3%) and standard amounts (Mann Whitney $P = 0.8$). Fifth, the median read depth of matched normal DNA (351x, IQR 139-756x) was lower than that of either primary tumour DNA sequenced at standard amounts (899x, IQR 527-1377x) or cfDNA sequenced at minute amounts (891x, IQR 530-1385x) (Mann Whitney $P = 0.001$ and $P < 0.001$). The median read depth of either cfDNA or primary tumour DNA sequenced at minute amounts was comparable to primary tumour DNA sequenced at standard amounts (Mann Whitney $P = 0.6$ and $P = 0.1$, respectively). Finally, after mapping of the sequencing reads, we called single nucleotide variants deviant from the reference genome to reveal germline single nucleotide polymorphisms (SNPs) and somatically acquired tumour-specific variants. Of

the called variants, the variant frequency was calculated as the fraction of variant reads over the total reads at that genomic position. Based on our criterion of 10 or more variant reads, matched normal material – intended to detect germline SNPs at 50% or 100% variant frequency – required a read depth of at least 20x. The percentage of bases at a depth of at least 20x was median 94.6% (IQR 91.8%-97.1%) in the matched normal DNA. Other material – intended to detect tumour specific somatic variants – required a read depth of at least 100x. In the primary tumour DNA and cfDNA the percentage of bases at a depth of at least 100x was respectively median 95.4% (IQR 95.0%-95.9%) and 95.0% (IQR 93.8%-95.8%).

Taken together, these findings imply that DNA sequenced at standard or at minute amounts results in comparable amounts and quality of data, but that sequencing material derived from FFPE preserved specimens – as is our matched normal material – resulted in less read depth and biased data with respect to length and GC content of the reads.

Defining the somatic origin of variants

To discriminate between SNPs and somatic variants in the primary tumour and cfDNA, we had FFPE-preserved tumour-adjacent normal mammary epithelium available for six of the ten patients. After calling single nucleotide variants deviant from the reference genome, an unexpected high and variable number of variants was detected in the matched normal DNA: a median of 1754 variants (IQR 78-4311 variants), as opposed to a median of 50 variants (IQR 46-54 variants) in the six corresponding primary tumour DNAs. The variants detected in matched normal DNA showed an enrichment for C > T transitions in their substitution spectrum, comprising median 86.0% (IQR 57.7-96.5%) of the total detected variants within a sample (**Figure 1A**, red). This is different than the distribution for SNPs in the human exome (i.e. 6.7% C > A, 8.9% C > G, 52.2% C > T, 4.4% T > A, 24.4% T > C and 3.5% T > G [4]) or that of the FF primary tumour material (**Supplementary Figure S2B**). Variants in matched normal DNA are only expected to represent hetero- and homozygous SNPs at variant frequencies of 50% and 100% respectively. The C > T transitions in these FFPE specimens often had variant frequency far below 50% in most samples (**Supplementary Figure S2A**). Therefore, we explored the effect of excluding from our matched normal FFPE-derived DNA all called variants which had variant frequencies below expected frequency for heterozygosity. Exclusion of variants below 35% variant frequency in these samples greatly reduced the number of reported variants to a median of 48 (IQR 41-297 variants) per sample. However, for P4 and P5, the number of variants detected remained unacceptable high after exclusion of variants below 35% variant frequency (resp. 379 and 543, **Figure 1B**) prompting us to omit all the data of those two matched normal samples for discrimination between

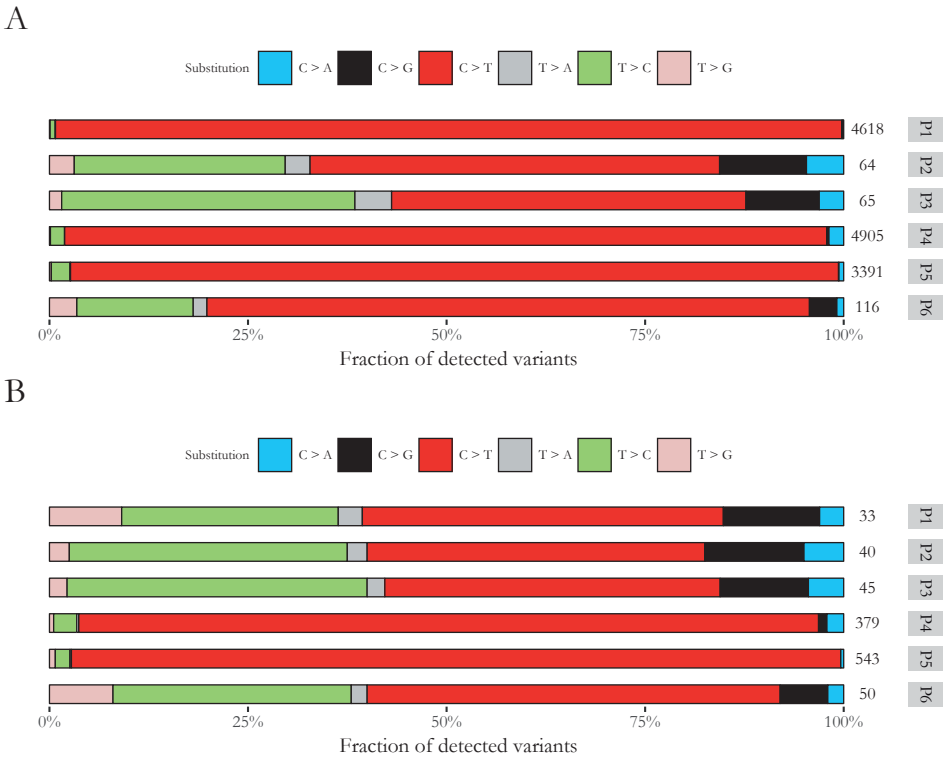


Figure 1 Substitution spectra of variants called relative to the reference genome in DNA derived from formalin fixed paraffin embedded (FFPE) matched normal specimens.
A: The contribution of the six possible base substitutions (C > A in blue, C > G in black, C > T in red, T > A in grey, T > C in green and T > G in pink) are depicted for each of six patients (P1 to P6) relative to the total variants in that sample. The total number of detected variants is depicted at the right end of the bars. **B:** A threshold removing variants with variant frequencies below 35% was applied to FFPE preserved matched normal specimens. The resulting substitution spectra (as in A) are depicted for each of the six patients (P1 to P6).

germline SNPs and somatic tumour variants. For the other four patients, the variants detected in matched normal DNA after ≤ 35% variant frequency curations were considered germline SNPs.

Next, we aimed to define the somatic or germline origin of variants detected in the primary tumour material of the ten patients. Alignment to the reference genome revealed a total of 486 alternative variants in all ten primary tumours with a median of 50 variants (IQR 45-54 variants) per individual tumour. The subsequent identification of somatically acquired variants in tumour material involves the exclusion of germline SNPs from the called variants. Conventionally, SNPs detected in matched normal DNA are used for this purpose (see above). Correction for germline SNPs reported in the curated matched normal DNA left us with 17, 6, 11 and 11 putative somatic variants in the primary

tumour of P1, P2, P3 and P6, respectively (**Table 1**). However, no suitable matched normal DNA for germline SNP detection was available for the remaining six patients. As an alternative approach, we explored the *in silico* Virtual Normal methodology [5]. This methodology applies the contextual information of variants reported in the public domain by not only taking into account the surrounding reference sequence, but also neighbouring variants. We compared the conventional method using matched normal DNA (MN) with the *in silico* method using Virtual Normal genomes (VN) for the four patients of which suitable matched normal material was sequenced (**Figure 2**). The concordance in classification as germline SNPs i.e. present in both the MN and VN, or as somatic variants i.e. absent from both the MN and VN, is in total median 81% of the variants detected in the primary tumours (**Figure 2**, black and blue). Absent from the MN but present in the VN were 15% of the variants (**Figure 2**, yellow) and 3% of the variants were absent from the VN but present in the MN (**Figure 2**, grey).

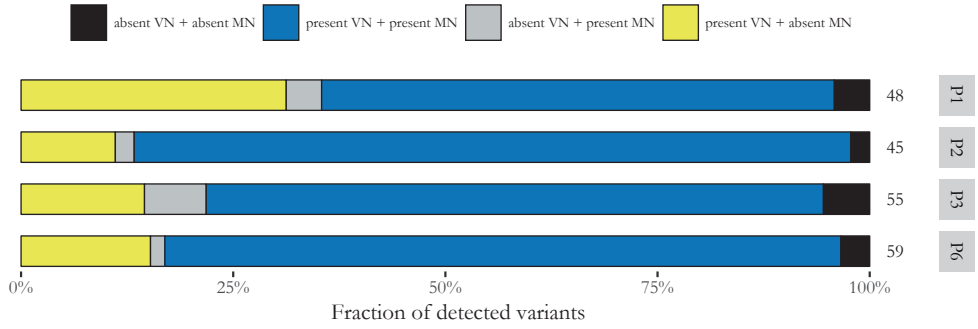


Figure 2 Defining somatic origin of variants using matched normal DNA (MN) or virtual normal genomes (VN) as reference.

Annotation of variants detected in the primary tumours of the four patients of who matched normal was accessible (P1, P2, P3 and P6) are depicted relative to the total variants in that sample. In here, variants absent from both the VN and MN in black, variants present in both VN and MN in blue, variants absent in the VN but present in the MN in grey, and variants present in the VN but absent from the MN in yellow. The total number of detected variants is depicted at the right end of the bars.

Taken together, we defined putative somatic variants in tumour DNA as follows. If matched normal material was available, occurrence of a tumour variant in this MN identifies variants as germline SNP. Additionally, irrespective the availability of matched normal material, *in silico* annotation using the VN genomes classifies variants as SNP when present in at least one VN genome. Applying the above to all variants of the ten primary tumours left us with 33 putative somatic variants recurrent at 26 distinct locations in the genome. Per individual tumour, a median of 2 (IQR 2-3) putative somatic variants were detected (**Table 1**, **Supplementary Table S1**).

Table 1 Variant detection in primary tumour and serum-derived cfDNA specimens.

Patient	Specimen	Variants	Variants in MN	Variants in VN	Somatic variants	Confirmed somatic variants
P1	Primary tumour	48	31	44	2	1
	cfDNA (serum T1)	46	29	42	2	2
	cfDNA (serum T2)	53	29	46	5	2
	cfDNA (serum T3)	77	30	47	28	1
P2	Primary tumour	45	39	43	1	1
	cfDNA (serum T1)	49	40	45	3	1
	cfDNA (serum T2)	46	40	44	1	1
	cfDNA (serum T3)	47	39	44	2	1
P3	Primary tumour	55	44	48	3	3
	cfDNA (serum T1)	57	45	49	4	3
	cfDNA (serum T2)	56	43	48	4	3
	cfDNA (serum T3)	55	44	48	3	3
P4	Primary tumour	36	NA	34	2	1
	cfDNA (serum T1)	33	NA	32	1	1
	cfDNA (serum T2)	33	NA	32	1	1
	cfDNA (serum T3)	35	NA	34	1	1
P5	Primary tumour	52	NA	49	3	3
	cfDNA (serum T1)	50	NA	47	3	3
	cfDNA (serum T2)	51	NA	48	3	3
	cfDNA (serum T3)	51	NA	48	3	3
P6	Primary tumour	59	48	56	2	1
	cfDNA (serum T1)	58	49	56	1	1
	cfDNA (serum T2)	59	50	57	1	1
	cfDNA (serum T3)	58	49	56	1	1
P7	Primary tumour	45	NA	35	10	8
	cfDNA (serum T1)	45	NA	34	11	7
	cfDNA (serum T2)	50	NA	36	14	6
	cfDNA (serum T3)	43	NA	34	9	8
P8	Primary tumour	51	NA	45	6	3
	cfDNA (serum T1)	49	NA	45	4	4
	cfDNA (serum T2)	49	NA	45	4	4
	cfDNA (serum T3)	48	NA	45	3	3
P9	Primary tumour	54	NA	52	2	2
	cfDNA (serum T1)	54	NA	52	2	1
	cfDNA (serum T2)	55	NA	54	1	1
	cfDNA (serum T3)	70	NA	54	16	1
P10	Primary tumour	41	NA	39	2	1
	cfDNA (serum T1)	111	NA	42	69	0
	cfDNA (serum T2)	54	NA	40	14	0
	cfDNA (serum T3)	44	NA	39	5	1

Number of variants (columns) in each specimen for each of the ten patients (rows). The columns indicate 1) total variants detected, 2) variants in the indicated specimen also detected in the matched normal specimen (if available), 3) variants in the indicated specimen also present in at least one Virtual Normal genome, 4) variants classified as somatic variant (criteria in manuscript) and 5) somatic variants in the indicated specimen also confirmed in at least one additional patient-matched specimen.

Detection of variants in minute material

To evaluate the feasibility of detecting variants in minute quantities of DNA, we compared the sequence output of our targeted sequencing panel on primary tumour DNA as standard and minute sequencing input. To this end, we sequenced for two individual patients one standard as 10 ng input and two minute counterparts as 250 pg input. In the first patient, we detected a total of 48 variants in the primary tumour sequenced at standard amounts and 46 and 49 variants in the two replicates sequenced at minute amounts (P1, **Table 2**). In the second patient, we detected a total of 45 variants in the primary tumour sequenced at standard amounts and 79 and 50 variants in the two replicates sequenced at minute amounts (P2, **Table 2**). Using the variants detected in the standard primary tumour DNA as predicted positives, the variant detection in minute replicates has a sensitivity of 96.8% (IQR 3.1%) and a false discovery rate of 9% (IQR 14.4%). The high discovery rate implies false positive variants generated as a result of sequencing minute quantities. Indeed, a high number of variants is detected in only a single minute replicates sample and are present at a low variant frequency (**Figure 3A**). To overcome these false positives, we aimed at the confirmation of a variant in an additional sample, which means we require a variant to be detected in at least two samples of an individual patient. A total of 44 variants were detected in both minute replicate samples for both P1 and P2. When we apply this confirmation-approach, a sensitivity of 91.7% (P1) and 97.8% (P2), and a false discovery rate of 0% (P1 and P2) were obtained. We next explored if the additional PCR cycles for the minute quantities introduced biases, such as a shift in variant frequency or allelic dropouts. For this, we selected variants in the primary tumour

Table 2 Variant detection in primary tumour specimen replicates at standard or minute input.

Patient	Replicate	Variants	Variants in MN	Variants in VN	Somatic variants	Somatic variants in 1/3	Somatic variants in 2/3	Somatic variants in 3/3
P1	Standard	48	31	44	2	0	1	1
	Minute (A)	46	29	41	3	1	1	1
	Minute (B)	49	30	44	3	2	0	1
P2	Standard	45	39	43	1	0	0	1
	Minute (A)	79	39	46	32	31	0	1
	Minute (B)	50	38	45	5	4	0	1

Number of variants detected (columns) in each replicate for each of the two patients of primary tumour specimen sequenced at standard and minute amounts (rows). The columns indicate 1) total variants detected, 2) variants in the indicated replicate also detected in the matched normal (MN) specimen (if available), 3) variants in the indicated replicate also present in at least one Virtual Normal (VN) genome, 4) variants classified as somatic variant (criteria in manuscript), 5) somatic variants in only the indicated replicate (one out of three) 6) somatic variants in the indicated and in one additional replicate (two out of three) and 7) somatic variants in the indicated and in all additional replicates (two out of three).

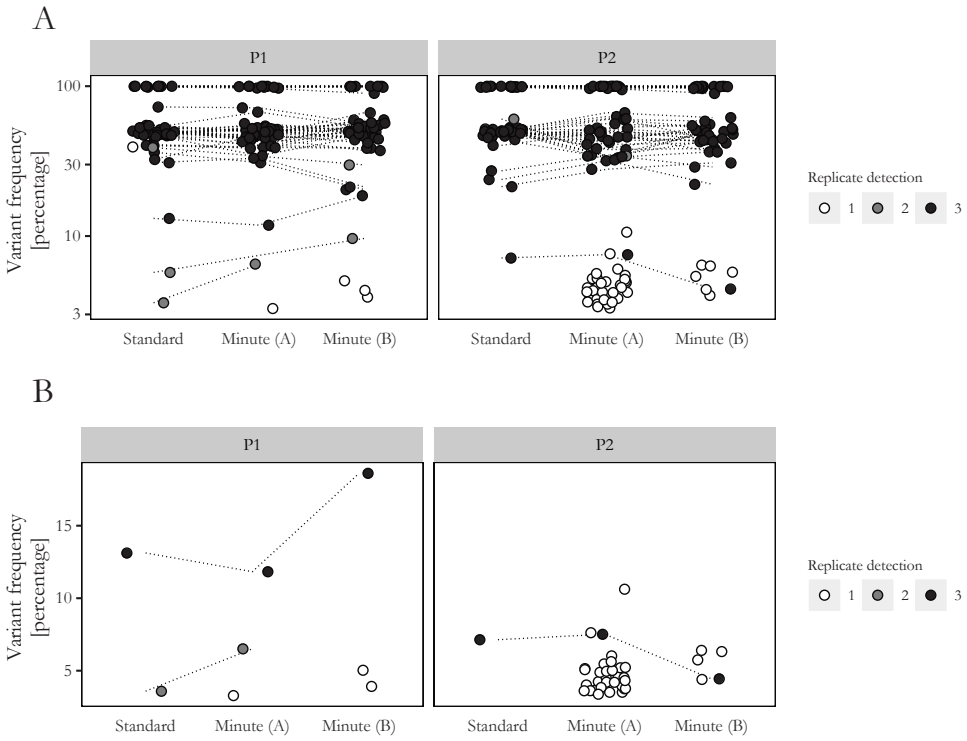


Figure 3 Detection of (somatic) variants in standard and minute primary tumour replicate samples.
A: The variant frequency in percentages of variants detected in a single replicate (1, white), in two out of three replicates (2, grey) or in all three replicates (3, black) of the two analysed patients. Recurrent somatic variants at unique genomic positions are connected by lines to visualize variant frequency between replicates. **B**: The variant frequency in percentages of somatic variants detected in a single replicate (1, white), in two out of three replicates (2, grey) or in all three replicates (3, black) of the two analysed patients. Recurrent somatic variants at unique genomic positions are connected by lines to visualize variant frequency between replicates.

defined as germline by their presence in the VN genomes and with a variant frequency between 45-55% (heterozygotes, n = 46) or above 95% (homozygotes, n = 26). In the primary tumours sequenced at standard quantities, the median variant frequency is for heterozygotes 49.4% (IQR 3.57%) and for homozygotes 99.7% (IQR 0.74%). In the minute replicates, a larger variation is visible for those selected variants: for heterozygotes a median variant frequency of 49.2% (IQR 8.7%) (Brown-Forsythe Levene-type test P < 0.001) and for homozygotes 99.7% (IQR 1.5%) (Brown-Forsythe Levene-type test P = 0.08). However, we did not observe any allelic dropouts (a shift in variant frequency from 45-55% to <5% or >95%, or from >95% to <5%) (**Supplementary Figure S3A**).

Last, for standard and minute replicates of the primary tumours, we defined the putative somatic variants for both patients (as described above). In the primary tumour

DNA sequenced at standard input we detected a total of 2 and 1 somatic variants in P1 and P2, respectively. Of those variants, we validated in the minute replicates 2/2 (100%) (Minute A) or 1/2 (50%) (Minute B) somatic variants in P1 and 1/1 (100%) (Minute A) or 1/1 (100%) (Minute B) somatic variants in P2 (**Figure 3B**). Thus, one of the three somatic variants was detected in one of the two minute replicates and therefore not detected by sequencing minute amounts.

Taken together, by sequencing only minute input quantities the majority of variants are detected however with a more variable variant frequency. The confirmation of variants in an additional patient-matched sample is necessary to minimize false discoveries, and this is therefore included in our workflow.

Sequencing minute cfDNA as tumour surrogate

Besides the primary tumours, we sequenced three consecutive serum-derived cfDNA samples at minute quantities for each of the ten patients, in which we defined the putative somatic variants (as described above). In the ten primary tumours we detected 33 putative somatic variants with median 2 (IQR 2-3) per individual, in the thirty sera we detected 219 putative somatic variants with median 3 (IQR 1-5) per cfDNA sample (**Table 1**). Because we included the confirmation of variants in an additional patient-matched sample, we first compared the putative somatic variants detected in each cfDNA sample with the matching primary tumour of each individual (**Figure 4**). A total of 24 putative somatic variants were present in at least one cfDNA sample and the matching primary tumour. On the other hand, there were 2 putative somatic variants confirmed in multiple matching cfDNA samples of an individual that were absent from the matching primary tumour (**Figure 4**, purple). Collectively, we were able to identify median 1 (IQR 1-3) putative somatic variants per cfDNA sample by confirmation in at least one additional patient-matched sample (**Table 1**). These 26 variants have potential as informative tumour-surrogate markers.

As a final step, we carefully inspected these variants to be able to use them as informative tumour surrogates (**Table 3**). First, we inspected the frequency of the variant. The contribution of non-tumour DNA is expected to be different between primary tumour specimen and cfDNA, and thus it would be unlikely for a somatic variant to have a cfDNA variant frequency similar to the variant frequency in the primary tumour specimen. Second, we employed public databases as *in silico* repositories for additional information about the detected variants, including databases on cancer driver genes [6], variants in relation to human health [7], germline polymorphisms [8-11] and functional consequences of the observed substitution [12]. Third, to make sure the variant had not been filtered out in other patient-matched samples, we inspected the detected somatic

Table 3 Specifics of confirmed somatic variants in primary tumour and serum-derived cfDNA specimens.

Patient	Specimen	Gene	Variant	Variant frequency	Public databases							Filtering	Informative tumour surrogate
					COSMIC	ClinVar	GoNL	dbSNP	ESP	1000G	RefSeq		
P1	Primary tumour	*AKAP9*	c.1686T>G	no call							Missense	No	Yes
	cfDNA (serum T1)			16.5%									
	cfDNA (serum T2)			17.0%									
	cfDNA (serum T3)			no call									
	Primary tumour	*NCOR1*	c.540G>C	13.1%							Missense	MN (8.4%)	No[†]
	cfDNA (serum T1)			10.3%									
	cfDNA (serum T2)			6.4%									
	cfDNA (serum T3)			14.2%									
P2	Primary tumour	*KMT2C*	c.1005T>A	7.1%	1 (other)			rs141993954	0.02		Silent	MN (6.5%)	No[†§]
	cfDNA (serum T1)			5.9%									
	cfDNA (serum T2)			7.4%									
	cfDNA (serum T3)			4.6%									
P3	Primary tumour	*KMT2C*	c.1005T>A	7.8%	1 (other)			rs141993954	0.02		Silent	MN (9%)	No[†§]
	cfDNA (serum T1)			7.8%									
	cfDNA (serum T2)			9.2%									
	cfDNA (serum T3)			9.3%									
	Primary tumour	*NCOR1*	c.540G>C	12.9%							Missense	MN (9.4%)	No[†]
	cfDNA (serum T1)			16.3%									
	cfDNA (serum T2)			15.5%									
	cfDNA (serum T3)			18.2%									
P3	Primary tumour	*PDE4DIP*	c.6933A>G	15.9%				rs3851872			Silent	MN (23.1%)	No[†§]
	cfDNA (serum T1)			24.4%									
	cfDNA (serum T2)			27.2%									
	cfDNA (serum T3)			22.1%									

P4	Primary tumour	*LRP2*	c.13685T>C	50.2%		0.01	rs142245618	0	0	Missense	MN (100%)	No‡§
	cfDNA (serum T1)			46.2%								
	cfDNA (serum T2)			51.3%								
	cfDNA (serum T3)			50.9%								
P5	Primary tumour	*AKAP9*	c.7541T>G	64.8%						Missense	MN (67.3%)	No‡
	cfDNA (serum T1)			49.5%								
	cfDNA (serum T2)			54.9%								
	cfDNA (serum T3)			46.5%								
	Primary tumour	*KMT2C*	c.1005T>A	6.0%	1 (other)		rs141993954	0.02		Silent	MN (9.2%)	No†§
	cfDNA (serum T1)			6.5%								
	cfDNA (serum T2)			7.9%								
	cfDNA (serum T3)			9.1%								
	Primary tumour	*NCOR1*	c.468A>G	5.5%						Silent	MN (6.2%)	No†
	cfDNA (serum T1)			9.7%								
	cfDNA (serum T2)			9.2%								
	cfDNA (serum T3)			8.4%								
P6	Primary tumour	*KMT2C*	c.1005T>A	6.9%	1 (other)		rs141993954	0.02		Silent	MN (6.4%)	No†§
	cfDNA (serum T1)			8.0%								
	cfDNA (serum T2)			7.7%								
	cfDNA (serum T3)			7.5%								
P7	Primary tumour	*APC*	c.7514G>A	47.5%		0	rs147549623	0		Missense	-	No†§
	cfDNA (serum T1)			34.7%								
	cfDNA (serum T2)			40.5%								
	cfDNA (serum T3)			50.1%								
	Primary tumour	*CDH1*	c.2336G>A	69.7%						Missense	-	No‡
	cfDNA (serum T1)			44.1%								
	cfDNA (serum T2)			36.5%								
	cfDNA (serum T3)			59.2%								

Table 3 continues on next page.

Table 3 *Continued.*

Patient	Specimen	Gene	Variant	Variant frequency	Public databases							Filtering	Informative tumour surrogate
					COSMIC	ClinVar	GoNL	dbSNP	ESP	1000G	RefSeq		
P7	Primary tumour	*KMT2C*	c.3384C>T	46.3%			0	rs144068847	0	0	Silent	-	No[‡§]
	cfDNA (serum T1)			41.1%									
	cfDNA (serum T2)			51.8%									
	cfDNA (serum T3)			50.6%									
	Primary tumour	*PDE4DIP*	c.6942T>C	37.9%				rs78461771			Silent	-	No[†§]
	cfDNA (serum T1)			28.0%									
	cfDNA (serum T2)			31.4%									
	cfDNA (serum T3)			36.7%									
	Primary tumour	*PIK3CA*	c.3140A>T	25.3%	>1 (breast)						Missense	S1 (2%)	Yes
	cfDNA (serum T1)			no call									
	cfDNA (serum T2)			no call									
	cfDNA (serum T3)			7.0%									
	Primary tumour	*RNF213*	c.10717G>A	61.6%							Missense	-	No[‡]
	cfDNA (serum T1)			35.6%									
	cfDNA (serum T2)			43.8%									
	cfDNA (serum T3)			55.0%									
	Primary tumour	*SMAD4*	c.1059C>A	17.0%							Nonsense	No	Yes
	cfDNA (serum T1)			8.5%									
	cfDNA (serum T2)			no call									
	cfDNA (serum T3)			4.3%									
	Primary tumour	*TP53*	c.639T>C	61.9%							Silent	-	No[‡]
	cfDNA (serum T1)			53.0%									
	cfDNA (serum T2)			54.6%									
	cfDNA (serum T3)			51.1%									
P8	Primary tumour	*APC*	c.7514G>A	35.8%			0	rs147549623	0		Missense	-	No[‡§]
	cfDNA (serum T1)			48.2%									
	cfDNA (serum T2)			41.4%									
	cfDNA (serum T3)			48.4%									

P8	Primary tumour	*NF1*	c.528T>A	4.5%	1 (other)	1	0.01	rs112306990	0	0	Missense	PT (34.4%) S3 (45.5%)	No†§
	cfDNA (serum T1)			6.9%									
	cfDNA (serum T2)			5.6%									
	cfDNA (serum T3)			6.8%									
	Primary tumour	*PDE4DIP*	c.2997C>T	no call					0		Silent	-	No†§
	cfDNA (serum T1)			55.8%									
	cfDNA (serum T2)			54.2%									
	cfDNA (serum T3)			no call									
P9	Primary tumour	*ARID1A*	c.4120C>T	23.5%							Silent	S1 (49.6%)	No‡
	cfDNA (serum T1)			21.4%									
	cfDNA (serum T2)			20.6%									
	cfDNA (serum T3)			17.6%									
	Primary tumour	*TP53*	c.520C>T	56.5%	>1 (breast)						Nonsense	No	Yes
	cfDNA (serum T1)			no call									
	cfDNA (serum T2)			60.3%									
	cfDNA (serum T3)			48.0%									
P10	Primary tumour	*CREBBP*	c.2728A>G	35.4%							Missense	S1 (33.1%) S2 (35.2%)	No‡§
	cfDNA (serum T1)			14.3%									
	cfDNA (serum T2)			no call									
	cfDNA (serum T3)			no call									
	Primary tumour			40.5%									
	cfDNA (serum T1)			no call									
	cfDNA (serum T2)			no call			0	rs143247685	0	0			
	cfDNA (serum T3)			62.5%									

Further inspection of the confirmed somatic variants in each specimen for each of the ten patients (rows). The columns indicate 1) targeted gene, 2) genomic location followed by the observed substitution as opposed to the reference genome, 3) in case of a call, variant frequency in indicated sample, 4) observation in public databases i.e. COSMIC, ClinVar, GoNL, dbSNP, ESP 1000G and RefSeq, 5) if the variant had been filtered out in patient-matched samples and if so, which samples with what variant frequency and 6) conclusion on the variants as an informative tumour surrogate with reasoning.

† Variant frequency similar in all patient-matched samples.

‡ Variant frequency at heterozygosity in all patient-matched samples.

§ Variant present in public polymorphism database.

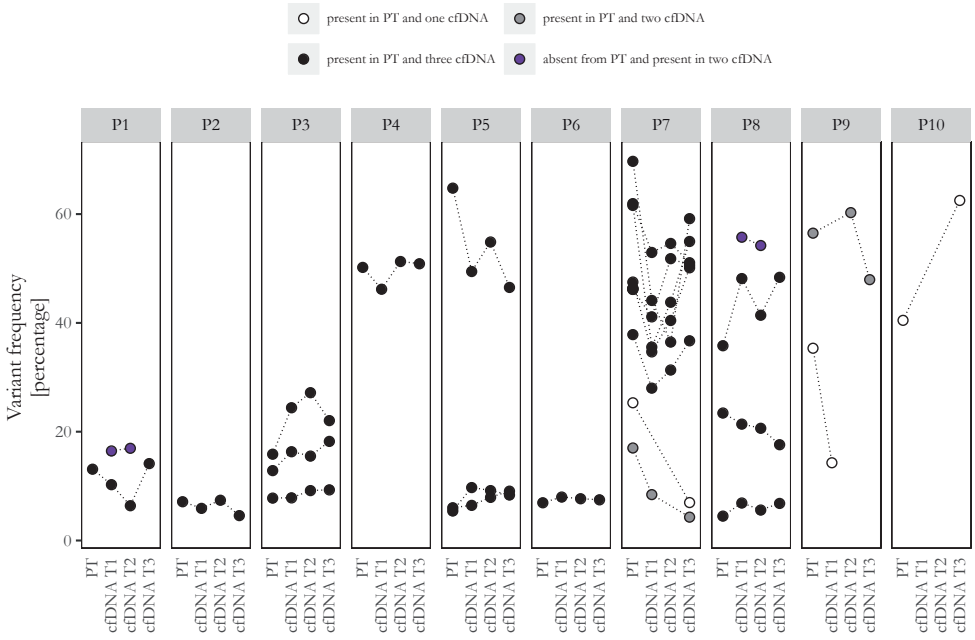


Figure 4 Detection of somatic variants in primary tumour and serum-derived cfDNA specimens.
The variant frequency in percentages of somatic variants detected in the primary tumour (PT) and three consecutive cfDNA samples (either T1, T2 or T3) of the ten patients in the study. Recurrent somatic variants at unique genomic positions are connected by lines to visualize variant frequency between replicates. Somatic variants detected in the primary tumour and confirmed in one, two or three patient-matched cfDNA sample in respectively white, grey or black. Somatic variants absent from the primary tumour but detected in two cfDNA sample in purple.

variants in the original variant call files without any of the applied filtering steps of all the patient-matched samples. By using these three criteria, 22 variants are considered as non-informative (**Table 3**). The remaining four variants – *AKAP9* c. 1686T>G in P1, *PIK3CA* c.3140A>T and *SMAD4* c.1059C>A in P7, *TP53* c.520C>T in P9 –are considered informative tumour surrogates traceable as ctDNA by our targeted sequencing approach (**Table 3**). These four variants were validated by conventional Sanger sequencing, and/or independent re-sequencing (**Supplementary Figure S4**).

Discussion

We aimed at developing an easily implementable targeted NGS approach to detect tumour-specific somatic variants in only minute amounts of cfDNA, to apply this to retrospective cohorts with specimens collected over time. These cohorts often have restrictions, such as limited amounts available and suboptimal preservation of specimens. To have a comprehensive coverage of somatic variants, we designed an amplicon panel

targeting exonic regions of 45 genes frequently mutated in breast, colon, prostate and ovarian cancer. Because of the small amplicon length in its design, this panel is suitable for the sequencing of fragmented DNA such as apoptotic cfDNA and degraded FFPE-derived DNA. Using this panel, we generated deep sequencing data of primary tumour DNA and minute amounts of serum-derived cfDNA samples of ten metastatic breast cancer patients from a retrospective cohort. We focused on the detection of single nucleotide substitutions in the sequencing data, because its detection is well-defined and it is the most prevalent mutation type in breast cancer [13].

The only source of germline DNA was FFPE-preserved tumour specimen from which we were able to obtain adjacent healthy tissue for six of the ten patients. We observed a variable number of artefacts in these FFPE-derived DNA samples. False discovery of particularly C > T transitions is known for FFPE-derived DNA samples [14-17], where cytosine residues are progressively de-aminated into uracil and upon amplification read as thymidine residues (C > U > T). This phenomenon is also evident in GC content of the generated reads, where for the matched normal specimens a decrease in GC content (i.e. C > T transition) is observed (**Supplementary Figure S1B**). The specimens were collected between 25 and 37 years ago, which may partially explain the suboptimal quality of these samples. For the detection of germline SNPs, the detection threshold of at least 35% on variant frequency was sufficient to eliminate the majority of the preservation-induced artefacts such as the C > T substitutions in four of the six cases. When material is not intended to only call germline SNPs but also lower frequency variants, i.e. in FFPE preserved tumour specimens, alternative strategies to eliminate FFPE-induced deamination products prior to sequencing – such as enzymatic treatment of the DNA samples or the use of proof-reading polymerases – are likely required [15, 17, 18].

As a next step, we aimed to define somatically acquired tumour-specific variants. The conventional approach uses matched normal DNA to make the distinction between somatic variants and germline SNPs in tumour material. However, when matched normal DNA is not of optimal quality – such as long-term FFPE preservation – the discrimination between germline SNPs and somatic variants becomes challenging. When artefacts are present in the matched normal data, there is a minor risk of erroneous categorization of tumour-specific variants as SNPs in the tumour material. Especially since the C > T transition is also often observed in somatic mutation profiles of cancer specimens [19]. Also, there is a major risk of classifying a germline SNP as somatic variant in tumour material when it is not detected in matched normal DNA. This seems to occur frequently for P1 (**Figure 2**): the addition of an alternative approach to discriminate between somatic and germline variants by using VN genomes [5] results in a large fraction of variants absent from the matched normal but present in the VN genomes. Also the fraction of variants present in both the matched normal and the VN genomes is small for P1. This indicates that some

germline variants had been missed in the poor quality matched normal DNA. The *in silico* approach using VN genomes thus removed additional variants which had been missed by taking only the matched normal approach. By combining our matched normal material and VN genomes, we identified the putative somatic variants in the tumour genomes of the ten patients described here. As an additional control for our classification as germline or somatic variant, we reason that tumour cell content of primary tumour specimen or the tumour-derived cfDNA fraction is never 100%, and thus that somatic variants cannot display a variant frequency close to 100%. Indeed, none of the variants with a variant frequency above 95% were classified as somatic (**Supplementary Figure S5A and S5B**).

By sequencing only minute quantities, we observed in a few samples some unexpected high numbers of variants i.e. in a primary tumour sample sequenced at minute amounts (P2 Minute A, **Table 2**) and also a few serum-derived cfDNA samples sequenced at minute amounts (P1 serum T3, P9 serum T3, P10 serum T1, **Table 1**). Theoretically, polymerase introduced artefacts – with a mutation rate of high-fidelity polymerases in the order of one mutation per a million bases – can reach approximately 300 variants per targeted amplification (worst-case scenario with 21 PCR cycle amplification of a 139 bp amplicon). However, these randomly introduced artefacts should not exceed variant frequencies above 1.5% for minute quantities (a single artefact in 250 pg representing 36 diploid cells [20]) and 0.05% for standard quantities (a single artefact in 10 ng representing 1429 diploid cells), which is both below the 2% threshold used by the variant caller. However, we did observe a larger variation in variant frequency when using only minute input amounts for the primary tumour minute replicates (**Supplementary Figure S3A**) and also for the cfDNA samples (**Supplementary Figure S3B**). Amplification bias and variability in length of the fragmented DNA might skew the amplification of specific DNA fragments and thus variant frequency of detected variants [21]. We were unable to discover the exact reason why an increase in detected variants occurred in a few samples, and thus confirmation of detected variants is necessary. Commonly, validation of variants is achieved by independent re-analysis of the DNA sample. However, in retrospective cohorts this is not always possible due to restrictive sources of DNA. As an alternative, we suggest the confirmation of variants through the detection in at least two DNA samples of an individual patient. Because our cfDNA samples originate from consecutive serum draws, this means we might miss variants such as those occurring only during a specific stage in the disease trajectory. Also, we are aware that an obvious drawback of using only minute quantities of cfDNA is false negative detection. We can expect variants – especially low frequent variants – to be missed just by chance because of using minute fractions representing forty times fewer molecules than the standard input.

Thus, we were able to detect multiple putative somatic variants in primary tumour material as well as their corresponding serum-derived cfDNA samples. As a final step,

we further inspected these variants whether they truly represent informative tumour surrogates. First, we compared variant frequency between patient-matched samples. The fraction of tumour-derived cfDNA is expected to be only a few percentages [22], whereas the primary tumour specimens are expected to have high tumour cell content. Therefore, somatic variants at around 50% variant frequency in both the primary tumour as well as the cfDNA are likely heterozygous SNPs; somatic variants at a low variant frequency in both the primary tumour as well as the cfDNA are potential artefacts. Somatic variants with a clearly different variant frequency between the primary tumour and cfDNA, or between consecutive cfDNA draws at different disease stages (e.g. disease recurrence), are indicative for true somatic variants traceable as ctDNA. Second, we took an advantage of public databases on variants. However, we observe conflicting annotation, i.e. variants being reported as both germline SNP but also as a tumour-specific somatic mutation in public databases (**Table 3**). Thus, this information should be taken with caution. Third, we verified if – due to stringent threshold settings – the detected variants had not been filtered out in patient-matched samples. Based on these three aspects we defined if the variant detected in cfDNA are to be used as an informative tumour surrogate (**Table 3**).

A limitation by using only minute quantities of cfDNA is the detection limit in the order of 2% variant frequency (250 pg represents approximately 36 diploid cells [20] thus allowing the detection of 1 mutant copy in 72 wildtype copies), whereas tumour-specific cfDNA can be present at variant frequencies as low as 0.001% [22]. Also, the retrospective serum samples used in this study are likely contaminated with DNA from non-tumour cells such as lysed leukocytes [23], and thus the fraction of tumour-derived cfDNA is expected to be low. We are aware that the preservative in the blood collection tube (i.e. cloth activator or anti-coagulant), and processing procedure of the blood-derivative (i.e. time-to-processing and sedimentation speed), has a great influence on the cfDNA quantity [24-27], and because this affects the fraction of tumour-derived cfDNA it should be kept in mind when analysing cfDNA. Also, this makes comparison with other findings in literature quite complex. Previous studies have assessed cfDNA in metastatic breast cancer using methods varying from multi-gene to single mutation approaches including targeted or exome NGS, Sanger sequencing, digital (droplet) PCR, and BEAMing, most often in plasma samples. In comparison, our detection rate of 30% (3/10 patients with tumour-specific cfDNA) is at the low side compared to the reported detection rate of 50%-100% in studies with a similar approach (profiling multiple genes in both the primary tumour and the cfDNA [28-32]). We attribute this difference to the use of minute quantities and thus our ability to detect variants at or above 2% VAF, since in the above mentioned studies approximately 30% of the reported tumour-specific variants in metastatic breast cancer have a frequency below 2%. In addition, we designed our gene-panel for the most frequently mutated genes, but variants not included in the panel will not be detected.

Taken together, in a retrospective cohort with only limited amounts of serum-derived cfDNA available, using our developed workflow, we were able to retrace in cfDNA somatic variants detected in the primary tumour and also somatic variants not detected in the primary tumour for three of the ten patients (30%). We conclude that the presented approach enables specific detection of tumour-specific somatic variants above 2% variant frequency in minute amounts of cfDNA and can be used to discover tumour surrogate markers to explore the potential of cfDNA in oncology.

Materials and methods

Specimens

Retrospectively collected specimens from metastatic breast cancer patients were obtained from our biobank. Selection was based on availability of fresh frozen primary tumour and serum samples at three different time points during treatment for metastatic disease and stored at -80°C. The consecutive serum samples were collected at start of first line tamoxifen therapy (T1), during therapy (T2) and at disease progression (T3). For some cases haematoxylin stained microscopic sections of the routine formalin fixed and paraffin embedded primary tumour was available, of which macro-dissected adjacent normal epithelial mammary tissue yielded matched normal specimen [33]. The study was approved by the medical ethics committee (MEC 02.953) and performed according to the Code of Conduct of Medical Scientific Societies (www.federa.org/codes-conduct). In the Netherlands, according to the Code of Conduct, informed consent is not required for retrospective analysis of bio specimens retrieved during standard of care procedures.

DNA extraction

Primary tumour DNA was extracted using the phenol chloroform method from pulverized FF specimens [34]. Matched normal DNA was extracted using Chelex 100 resin (*Bio-Rad, Veenendaal, the Netherlands*) as described previously [33]. CfDNA was extracted after external lysis of the serum (400 µL) using the automated MagNA Pure Compact Nucleic Acid Isolation Kit I (*Roche Diagnostics, Almere, the Netherlands*) [35]. Extracted DNA was quantified using the Qubit® 2.0 fluorometer (*Thermo Fisher Scientific, Landsmeer, the Netherlands*).

Amplicon-based targeted next-generation sequencing

Ion semiconductor sequencing on the Ion Torrent Personal Genome Machine (PGM) was performed with an Ion AmpliSeq custom Panel applying consumables, kits, software

packages and protocols of the manufacturer (*Thermo Fisher Scientific*). In short, adapter-ligated libraries were constructed using the AmpliSeq Library kit 2.0 using 3106 amplicons designed for small regions (63 to 139 nucleotides insert size) targeting 45 cancer-related genes [35]. Gene selection was based on the most frequently mutated driver genes in breast, colon, prostate and ovarian cancer as revealed by extensive genomic analysis deposited in the COSMIC database and interrogates for 39 genes all coding exons and for 6 genes only those exons harbouring hotspot mutations (**Supplementary Table S2**), covering a total of 14159 COSMIC mutations. For sequencing of minute amounts, the recommended DNA starting input of 10 ng was reduced to approximately 250 pg (range 165.6-573.6 pg). For samples below 10 ng DNA input, adapter-ligated library preparation was adjusted from the standard 17 PCR cycles to 20 or 21 PCR cycles. After purification with AMPure XP beads (*Beckman Coulter, Woerden, the Netherlands*), library quantity was determined using the Ion Library Quantitation kit and diluted to a final concentration of 8 pM. Template was emulsion PCR-prepared using the Ion PGM Template OT2 200 kit on the Ion OneTouch 2 system and confirmed using the Ion Sphere quality control kit. Template-positive Ion Sphere Particles were enriched using DynaBeads MyOne Streptavidin C1 on the Ion OneTouch ES instrument and barcoded samples were sequenced on the Ion Torrent PGM for 500 flows using the Ion PGM Sequencing Kit v2.0 on an Ion 318v2 chip.

Bioinformatics and statistics

Data from the PGM runs were processed initially using the Ion Torrent platform-specific pipeline software Torrent Suite to generate sequence reads, trim adapter sequences, filter, and remove poor signal-profile reads. Quality control of generated reads was performed using the Torrent Suite Software v4.0 with the "Coverage Analysis" plug-in and using FastQC [36]. Initial variant calling compared to the reference genome hg19 (build 37) was generated using Torrent Suite Software v4.0 with the "variant caller v4.0" plug-in. To be able to detect low frequency variants with minimal false negative calls we used the Somatic - Low Stringency Optimized settings. To correct for panel-specific sequencing errors, we removed all variants detected in >90% of the analysed samples. Further analyses were conducted on single nucleotide variants in exonic regions only, a variant read depth of at least 10x, a quality score of at least 20 and with a minimal strand bias threshold of 0.9. We applied to matched normal material – intended to detect germline SNPs at 50% or 100% variant frequency – a criterion of a at least 20x read depth. Other material – intended to detect tumour specific somatic variants – required a read depth of at least 100x. The Virtual Normal methodology was applied using the Galaxy tool in the DTLS tool shed, using the 'Diversity and 1000G (479 genomes)' virtual normal set. Annotation

of the variants was performed by a custom pipeline including ANNOVAR [37], the Catalogue Of Somatic Mutations In Cancer [6] (COSMIC) version 67, ClinVar [7], the Genome of the Netherlands [11] (GoNL) version 7.0.0.59, Database of Single Nucleotide Polymorphisms [8] (dbSNP) build ID 137 (non-flagged), 1000 Genomes [10] (1000G) version 2012april_all, and the Exome Variant Server [9] (ESP) version 6500si_all. Variants assigned as somatic were examined visually using Integrative Genomics Viewer (IGV) software. Statistical comparisons were performed by the indicated tests in R version 3.2.3.

Supplementary data

Supplementary data for this article are available online at Scientific Reports (https://www.nature.com/srep/).

References

1. Fleischhacker, M. and B. Schmidt, *Circulating nucleic acids (CNAs) and cancer--a survey.* Biochim Biophys Acta, 2007. **1775**(1): p. 181-232.
2. Murtaza, M., et al., *Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA.* Nature, 2013. **497**(7447): p. 108-112.
3. Bettegowda, C., et al., *Detection of circulating tumor DNA in early- and late-stage human malignancies.* Sci Transl Med, 2014. **6**(224): 224ra24.
4. Jiang, C. and Z. Zhao, *Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms.* Genomics, 2006. **88**(5): p. 527-534.
5. Hiltemann, S., et al., *Discriminating somatic and germline mutations in tumor DNA samples without matching normals.* Genome Res, 2015. **25**(9): p. 1382-1390.
6. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(D1): D805-11.
7. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res, 2016. **44**(D1): D862-8.
8. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-311.
9. (ESP), N.G.E.S.P. *Exome Variant Server*. Available from: http://evs.gs.washington.edu/EVS/.
10. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
11. Genome of the Netherlands, C., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet, 2014. **46**(8): p. 818-825.
12. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): D733-45.
13. Stephens, P.J., et al., *The landscape of cancer genes and mutational processes in breast cancer.* Nature, 2012. **486**(7403): p. 400-404.
14. Do, H. and A. Dobrovic, *Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase.* Oncotarget, 2012. **3**(5): p. 546-558.
15. Do, H., et al., *Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates.* Clin Chem, 2013. **59**(9): p. 1376-1383.
16. Kerick, M., et al., *Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity.* BMC Med Genomics, 2011. **4**: 68.
17. Wong, S.Q., et al., *Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing.* BMC Med Genomics, 2014. **7**: 23.
18. Bourgon, R., et al., *High-throughput detection of clinically relevant mutations in archived tumor samples by multiplexed PCR and next-generation sequencing.* Clin Cancer Res, 2014. **20**(8): p. 2080-2091.
19. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-993.
20. Tiersch, T.R., et al., *Reference standards for flow cytometry and application in comparative studies of nuclear DNA content.* Cytometry, 1989. **10**(6): p. 706-710.
21. Kanagawa, T., *Bias and artifacts in multitemplate polymerase chain reactions (PCR).* J Biosci Bioeng, 2003. **96**(4): p. 317-323.
22. Canzoniero, J.V. and B.H. Park, *Use of cell free DNA in breast oncology.* Biochim Biophys Acta, 2016. **1865**(2): p. 266-274.
23. Chan, K.C., et al., *Effects of preanalytical factors on the molecular size of cell-free DNA in blood.* Clin Chem, 2005. **51**(4): p. 781-784.

24. Wong, F.C., et al., *Cell-free DNA in maternal plasma and serum: A comparison of quantity, quality and tissue origin using genomic and epigenomic approaches.* Clin Biochem, 2016. **49**(18): p. 1379-1386.
25. Parpart-Li, S., et al., *The effect of preservative and temperature on the analysis of circulating tumor DNA.* Clin Cancer Res, 2016. **23**(10): p. 2471-2477.
26. Medina Diaz, I., et al., *Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing.* PLoS One, 2016. **11**(11): e0166354.
27. van Dessel, L.F., et al., *Application of circulating tumor DNA in prospective clinical oncology trials: standardization of pre-analytical conditions.* Mol Oncol, 2017.
28. Dawson, S.J., et al., *Analysis of circulating tumor DNA to monitor metastatic breast cancer.* N Engl J Med, 2013. **368**(13): p. 1199-209.
29. Parsons, H.A., et al., *Individualized Molecular Analyses Guide Efforts (IMAGE): A Prospective Study of Molecular Profiling of Tissue and Blood in Metastatic Triple-Negative Breast Cancer.* Clin Cancer Res, 2017. **23**(2): p. 379-386.
30. Shaw, J.A., et al., *Mutation Analysis of Cell-Free DNA and Single Circulating Tumor Cells in Metastatic Breast Cancer Patients with High Circulating Tumor Cell Counts.* Clin Cancer Res, 2017. **23**(1): p. 88-96.
31. Page, K., et al., *Next Generation Sequencing of Circulating Cell-Free DNA for Evaluating Mutations and Gene Amplification in Metastatic Breast Cancer.* Clin Chem, 2017. **63**(2): p. 532-541.
32. Rothe, F., et al., *Plasma circulating tumor DNA as an alternative to metastatic biopsies for mutational analysis in breast cancer.* Ann Oncol, 2014. **25**(10): p. 1959-1965.
33. van Lier, M.G., et al., *A review on the molecular diagnostics of Lynch syndrome: a central role for the pathology laboratory.* J Cell Mol Med, 2010. **14**(1-2): p. 181-197.
34. Berns, E.M., et al., *c-myc amplification is a better prognostic factor than HER2/neu amplification in primary breast cancer.* Cancer Res, 1992. **52**(5): p. 1107-1113.
35. Jansen, M.P., et al., *Cell-free DNA mutations as biomarkers in breast cancer patients receiving tamoxifen.* Oncotarget, 2016. **7**(28): p. 43412-43418.
36. Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data*. 6-6-14; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
37. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): e164.

CHAPTER 6


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Asp
Ser
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Leu

Sensitive detection of mitochondrial DNA variants for analysis of mitochondrial DNA-enriched extracts from frozen tumour tissue

Marjolein J.A. Weerts | Eveline C. Timmermans | Rolf H.A.M. Vossen | Dianne van Strijp | Mirjam C.G.N. Van den Hout-van Vroonhoven | Wilfred F.J. van IJcken | Pieter-Jan van der Zaag | Seyed Y. Anvar | Stefan Sleijfer | John W.M. Martens

Scientific Reports 2018; 8:2261

Abstract

Large variation exists in mitochondrial DNA (mtDNA) not only between but also within individuals. Also in human cancer, tumour-specific mtDNA variation exists. In this work, we describe the comparison of four methods to extract mtDNA as pure as possible from frozen tumour tissue. Also, three state-of-the-art methods for sensitive detection of mtDNA variants were evaluated. The main aim was to develop a procedure to detect low-frequent single-nucleotide mtDNA-specific variants in frozen tumour tissue. We show that of the methods evaluated, DNA extracted from cytosol fractions following exonuclease treatment results in highest mtDNA yield and purity from frozen tumour tissue (270-fold mtDNA enrichment). Next, we demonstrate the sensitivity of detection of low-frequent single-nucleotide mtDNA variants (≤ 1% allele frequency) in breast cancer cell lines MDA-MB-231 and MCF-7 by single-molecule real-time (SMRT) sequencing, UltraSEEK chemistry based mass spectrometry, and digital PCR. We also show *de novo* detection and allelic phasing of variants by SMRT sequencing. We conclude that our sensitive procedure to detect low-frequent single-nucleotide mtDNA variants from frozen tumour tissue is based on extraction of DNA from cytosol fractions followed by exonuclease treatment to obtain high mtDNA purity, and subsequent SMRT sequencing for (*de novo*) detection and allelic phasing of variants.

Introduction

The past decades, extensive genomic analysis of tumour specimens using massive parallel sequencing by large sequencing consortia (e.g. https://www.icgc.org/icgc and http://cancergenome.nih.gov/) have revealed the major somatic drivers of human cancer, that have been reported in numerous studies. However, the small circular genome of the mitochondria has been largely ignored in such analyses. The human mitochondrial DNA (mtDNA) consists of ~16,569 base pairs encoding 37 genes: two ribosomal RNAs and twenty-two transfer RNAs functioning in the mitochondrial translation apparatus and thirteen proteins essential for oxidative phosphorylation. The total number of mtDNA molecules per cell varies between cell types from a few up to several thousand, and depends on both the number of mitochondria per cell as well as the number of mtDNA molecules per mitochondrion [1-3]. Similar to chromosomal DNA in the nucleus (nDNA), mtDNA may contain rare or polymorphic variants. Currently nearly 10,000 variable positions within mtDNA are reported in public databases [4]. When variation is acquired, genetically different mtDNA molecules can reside within a single cell, referred to as heteroplasmy (that is, > 0% and < 100% allele frequency per cell). Importantly, heteroplasmic patterns can differ within an individual across tissues [5-8]. Despite inherited and somatically acquired variants in mtDNA being associated with multiple human diseases [9], the exact significance of somatic mtDNA variants in cancer remains controversial [10, 11].

Recently, taking advantage of publically available data from the large sequencing consortia, a handful of papers reported on the catalogue of somatic mitochondrial variants in multiple tumour types [12-14]. However, a complicating issue in the genomic analysis of mtDNA is the presence of sequences of mitochondrial origin in the nDNA (termed nuclear insertions of mitochondrial origin, NUMTs). NUMTs have likely originated from joining mtDNA/RNA fragments to nDNA ends during double strand break repair [15, 16] and are found in nearly all eukaryotes that contain mtDNA. This process may occur at any moment during lifetime [17] as well as during tumour evolution [18]. There are fixed NUMTs present in virtually every human genome – and thus reported in the human reference genome – inserted millions of years ago, but also more recent NUMT insertions have been described [19]. Unfortunately, due to their sequence similarity to mtDNA, NUMTs can interfere with accurate variant detection and thus investigation of mitochondrial heteroplasmy [16, 19-23]. Estimations based on the human reference genome indicate that for each 175 base pairs mtDNA segment an average of 9.5 NUMT copies are present in the human nDNA [24], but this number may likely be higher [19]. In addition, since the insertion of the mitochondrial genome is an ongoing process, this number is even larger in tumour cells since they also contain all somatic insertions events

of NUMTs [18]. In addition, in tumour cells the processes shaping nDNA [25, 26] are substantially different from the one that shapes the mtDNA [13], resulting in somatic variants in NUMTs and complicating accurate mtDNA heteroplasmy detection even further for tumour cells.

Consequently, the large variation in mtDNA between and within individuals as well as the presence of NUMTs demands a highly specific and sensitive detection of mtDNA variants, especially for low-frequent tumour-specific variants. In the study described here, we aimed to develop a sensitive procedure to detect low-frequent single-nucleotide mtDNA variants in frozen tumour tissue. Multiple efforts in developing methods for extraction of pure mtDNA exist [27-34]. These include methods using commercial kits or (laborious) ultracentrifugation to obtain pure mitochondria, and techniques to enrich for mtDNA by either the isolation technique or enzymatic degradation of nDNA. Unfortunately, the majority of previous studies focused on either cultured cells or cells from the blood and not on more physically and biochemically complex structures formed by tissue specimens. Thus, the application of these techniques to frozen tumour tissue specimens – an important source to assess tumour cell characteristics – has not been shown to date. Therefore, we compared four easily implementable procedures to extract mtDNA as pure as possible from frozen tumour tissue. Also, we evaluated three state-of-the-art techniques for the detection of low-frequent mtDNA-specific variants: Pacific Biosciences' SMRT sequencing [35], UltraSEEK chemistry [36] and digital PCR.

Results

Procedure to obtain mtDNA-enriched DNA extracts from frozen tumour tissue

To obtain mtDNA as pure as possible from frozen tumour tissue, our first focus was on the most optimal isolation procedure to extract mtDNA with minimal carry-over of nDNA. For this, we extracted DNA from fresh frozen primary tumour specimens using four easily implementable methods, and compared the yields via quantification of the percentage of mtDNA (**Figure 1A**) and total amount of dsDNA (**Figure 1B**). A silica-based total cellular DNA extraction method (I) used as reference for yield resulted in median 863 ng (interquartile range IQR 94 ng) dsDNA of which 0.1% (IQR 0.0%) mtDNA. A method (II) based on alkaline extraction – commonly used to extract plasmid DNA and thus designed to extract circular DNA [28, 30, 32, 33] – yielded median 144 ng (IQR 140 ng) dsDNA with 0.5% (IQR 0.6%) mtDNA. Extracting DNA from isolated mitochondria (III) [34] yielded median 825 ng (IQR 529 ng) dsDNA with 0.2% (IQR 0.1%) mtDNA. A selective lysis method (IV) that starts with the disruption of the plasma membrane to release the cellular components [29, 37] followed by sedimentation of cell nuclei, and

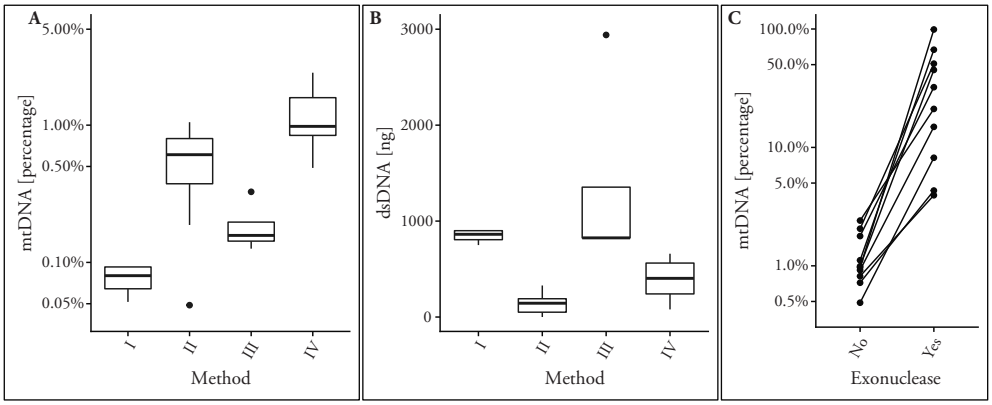


Figure 1 Comparison of methods for mtDNA extraction from fresh frozen tumour specimens.
Cryosections originating from ten frozen tumour specimens (biological replicates) were subjected to different extraction procedures including (I) a total cellular DNA extraction method, (II) a method based on alkaline extraction, (III) a method extracting DNA from isolated mitochondria and (IV) a selective lysis method extracting DNA from cytosol fractions. For each method, the percentage of mtDNA (A) and total amount of dsDNA (B) was quantified. Also, DNA extracts from cytosol fractions originating from ten frozen tumour specimens were subjected to exonuclease-based enrichment and the percentage of mtDNA quantified, with for each specimen the mtDNA percentage before and after treatment connected by lines (C). Boxplots represent median, inter quartile range (IQR) and 1.5 × IQR.

DNA extracted from the remaining cytosol fraction yielded median 403 ng (IQR 321 ng) dsDNA with 1.0% (IQR 0.8%) mtDNA. Note that a similar trend was obtained by these methods using frozen cultured cells as input (**Supplementary Figure 1A/B**). From these results, it is evident that the best isolation procedure to extract mtDNA from frozen tumour tissue is method IV – DNA from cytosol fractions – with the highest mtDNA percentage and sufficient dsDNA yield. To increase the mtDNA fraction, we applied an enzymatic exonuclease reaction to degrade specifically linear nDNA. This greatly increased the percentage of mtDNA in DNA extracts from cytosol fractions, from median 1% (IQR 0.8%) to median 27% (IQR 40%) (**Figure 1C**). This result was also obtained when using DNA from frozen cultured cells as input material (**Supplementary Figure 1C**). Exonuclease treatment on total cellular DNA extracts increased the percentage of mtDNA as well, but not to the same extent as for DNA extracts from cytosol fractions, and total dsDNA yield was lower (**Supplementary Figure 2**). Concluding, the preferred procedure to obtain mtDNA as pure as possible from fresh frozen tumour tissue is to extract DNA from cytosol fractions followed by exonuclease treatment.

Approach for sequencing of mtDNA

Next we explored sequencing methods for the detection of mtDNA variants. First, whole genome sequencing-by-synthesis (SBS) was applied to total cellular DNA extracts (method

I) and DNA extracts from cytosol fractions (method IV), both without and with additional enrichment for mtDNA by exonuclease treatment. As expected, the cell line DNA extract from cytosol fraction treated with exonuclease yielded the highest percentage of aligned reads to mtDNA (86%), whereas the other methods yielded much lower percentages (< 25%) (**Supplementary Table 1**). The DNA extract from cytosol fraction treated with exonuclease derived from fresh frozen tumour tissue yielded a percentage of aligned reads to mtDNA in line with the PCR-based mtDNA percentage (respectively 12% and 10%). Thus, despite the relatively high fraction of 10% mtDNA, a major proportion of reads were derived from nuclear DNA. The observed spread in mtDNA percentage in exonuclease treated method IV extracts from frozen tumour tissue (**Figure 1C**) will therefore lead to a variable proportion of mtDNA reads using whole genome SBS. To circumvent this variability, we decided to explore a targeted approach for sequencing mtDNA.

For this, nine primer sets covering the complete mtDNA were evaluated for their specificity to mtDNA, as *in silico* BLAST search showed that the primers did not match to known NUMT sequences in the reference genome. Specificity of the nine primer sets was confirmed by the absence of PCR products in two mtDNA-depleted cell lines (**Supplementary Figure 3**), allowing mtDNA-specific sequencing of the nine amplicons using single-molecule real-time (SMRT) sequencing. This method is able to generate long reads, covering each amplicon in a single read. To obtain an estimate of sequencing output and to evaluate variants detected by the whole genome SBS and targeted SMRT sequencing approaches, we compared for the two approaches the sequencing output of MDA-MB-231 DNA extracts from cytosol fraction treated with exonuclease. Whole genome SBS generated a total of 800,504 reads of 100 nucleotides (of which 87% duplicated reads) and after alignment resulted in an evenly distributed coverage of median 201x (IQR 2, range 13 - 404). The 2,727 reads of 1,738 - 2,836 base pairs by targeted SMRT sequencing displayed more variable coverage among the amplicons with median 282x (IQR 132, range 87 - 761) (**Supplementary Figure 4**). The more variable coverage in targeted SMRT sequencing was mainly due to regions where amplicons overlapped, causing an increase in coverage (**Supplementary Figure 4**). Both sequencing approaches detected all 29 positions with a documented alternative allele in MDA-MB-231 against rCRS at homoplasmic levels (> 99% allele frequency). Also additional heteroplasmic variants were detected, with no major differences observed between the two sequencing approaches (**Supplementary File**). Given the lower output in read depth per number of generated reads by whole genome SBS sequencing – due to a loss of reads which map to the nuclear genome – and the risk of introducing NUMTs hampering downstream analysis, we continued sequencing experiments using the targeted SMRT sequencing approach.

Sensitive detection of low-frequent mtDNA variants

To detect low-frequent single-nucleotide variants in mtDNA, we evaluated three approaches: SMRT sequencing, UltraSEEK chemistry and digital PCR. As a source of mtDNA we used breast cancer cell lines MDA-MB-231 and MCF-7. A total of respectively 29 and 13 variants alternative to rCRS have been documented in the mtDNA of MDA-MB-231 (also see above) and MCF-7, with a total of 28 positions containing a different allele between the two cell lines. To determine detection limits empirically, we prepared mixtures of the cell lines – considering MDA-MB-231 as the mutant variant – to generate samples with allele frequencies of 0%, 0.001%, 0.01%, 0.1%, 1% and 10% variant. The mixture samples were subjected to the three detection methods, and we evaluated their ability to detect the mutant variant. By SMRT sequencing, we obtained a median coverage of 4,060x per sample (IQR 4,842x, range 648 – 34,263x) (see **Supplementary**

Table 1 Limit of detection for low-frequent mtDNA variants by SMRT sequencing, UltraSEEK and digital PCR.

Position	Mutant (MDA-MB-231)	Wildtype (MCF-7)	Limit of detection per method		
			SMRT	UltraSEEK	Digital PCR
153	G	A	≥0.1%	na	na
195	C	T	≥0.1%	na	na
1719	A	G	≥0.1%	na	na
2706	G	A	≥0.1%	na	na
6221	C	T	≥1.0%	≥1.0%	na
6371	T	C	≥1.0%	≥0.1%*	na
6776	T	C	≥0.1%	na	na
7028	T	C	≥0.1%	na	na
8506	C	T	≥0.1%	≥1.0%	na
9966	G	A	≥0.1%	na	na
11719	A	G	≥0.1%	na	na
12084	T	C	≥0.1%	≥0.1%	na
12705	T	C	≥0.1%	na	na
13966	G	A	≥0.1%	≥0.1%	≥0.01%
14470	C	T	≥0.1%	≥0.1%	na
14766	T	C	≥0.1%	na	na
15310	C	T	≥0.1%	≥0.1%*	≥0.1%
15380	A	G	≥0.1%	na	na
16093	C	T	≥0.1%	na	na
16184	C	T	≥0.1%	na	na
16223	T	C	≥0.1%	na	na
16265	G	A	≥0.1%	na	na
16278	T	C	≥0.1%	na	na

Detection of the mutant variant allele (MDA-MB-231 genotype) in the lowest mutant fraction mixture indicated per position per method (empirical limit of detection). For the UltraSEEK and digital PCR method this was limited to respectively 7 and 2 positions due to requirement of generating dedicated PCR primers. na = not analysed. * = detected in 1 out of 3 replicate samples.

Table 2 Allele frequency of the heteroplasmic *de novo* variants detected in six cell line mixtures by SMRT sequencing.

Position	Variant	Detected amplicon[a]	Phased genotype[b]	Cell line mixture (mutant fraction)					
				0%	0.001%	0.01%	0.1%	1%	10%
76	T	A, B	Wildtype	24.06*	24.75*	24.44*	24.75*	24.35*	21.01*
15806	A	A	Wildtype	7.09	7.29	7.07	7.12	7.58	6.22
1062	A	B	Wildtype	1.29	1.34	1.24	1.33	1.20	1.37
10085	T	F	Wildtype	0.60	0.86	0.87	0.84	0.74	0.70
7029	T	E	Wildtype	0.49	0.48	nc	0.37	nc	0.47
14644	T	I	Wildtype	0.30	0.38	0.24	0.23	0.41	0.36
14817	T	I	Wildtype	0.29	0.23	0.33	0.34	nc	nc
72	C	A, B	Wildtype	0.13*	0.12*	0.14*	0.14*	0.19*	0.08*
15897	A	A	Wildtype	0.12	0.09	0.16	0.10	nc	0.15
1398	C	B	Wildtype	0.08	nc	0.08	0.05	nc	nc
39	T	A, B	Wildtype	0.06*	0.06*	0.03*	0.10*	0.08*	0.04*
5031	A	D	Wildtype	0.14	nc	nc	nc	0.17	nc
14751	T	I	Wildtype	nc	0.15	nc	0.09	0.16	nc
15129	C	I, A	Wildtype	nc	0.05*	0.05*	nc	nc	nc
934	A	B	Wildtype	nc	0.05	0.05	nc	nc	nc
564	A	B	Wildtype	nc	0.05	nc	0.08	nc	nc
12124	T	G, H	Wildtype	nc	0.05*	nc	0.04*	0.07*	nc
103	A	A, B	Wildtype	nc	nc	nc	0.01*	nc	nc
13680	T	H, I	Wildtype	nc	nc	nc	0.03*	nc	nc
10607	T	F, G	Wildtype	nc	nc	nc	nc	0.06*	nc
16391	A	A	Wildtype	0.07#	nc	nc	nc	nc	nc
9808	T	F	Wildtype	nc	nc	0.08#	nc	nc	nc
11778	A	G	Wildtype	nc	nc	nc	0.06#	nc	nc
14607	A	I	Wildtype	nc	nc	nc	0.06#	nc	nc
228	A	B	Wildtype	nc	nc	nc	0.05#	nc	nc
9627	A	F	Wildtype	nc	nc	nc	0.04#	nc	nc
9804	A	F	Wildtype	nc	nc	nc	0.04#	nc	nc
15550	T	A	Wildtype	nc	nc	nc	0.03#	nc	nc
15604	T	A	Wildtype	nc	nc	nc	0.03#	nc	nc
16067	T	A	Wildtype	nc	nc	nc	0.03#	nc	nc
16169	T	A	Wildtype	nc	nc	nc	0.03#	nc	nc
664	A	B	Wildtype	nc	nc	nc	nc	0.09#	nc
12818	A	H	Mutant	nc	nc	nc	0.06+	0.91+	8.03+
16184	A	A	Mutant	nc	nc	nc	0.07+	0.41+	6.89+
763	A	B	Mutant	nc	nc	nc	0.06+	0.61+	6.16+
13623	T	H, I	Mutant	nc	nc	nc	nc	nc	0.31*+
10406	A	F, G	Mutant	nc	nc	nc	nc	nc	0.18*
6887	T	E	Mutant	nc	nc	nc	nc	nc	0.88+
3714	G	C	Mutant	nc	nc	nc	nc	nc	0.34#
16218	T	A	Mutant	nc	nc	nc	nc	nc	0.22#
3697	A	C	Mutant	nc	nc	nc	nc	nc	0.23#
1323	A	B	Mutant	nc	nc	nc	nc	nc	0.14#

a = the amplicon (termed A to I) in which the variant was detected, which can be either one or two in the case of overlapping regions. b = the genotype of the variant as determined by allelic phasing (i.e. either MCF-7 considered wildtype or MDA-MB-231 considered mutant). nc = not called. * = variants that are detected in two overlapping amplicons and thus by two independent observations. + = variants that were detected in a sample containing 100% mutant material by both SMRT and SBS sequencing at a lower depth (Supplementary File). # = variants that can in theory be PCR errors because they were detected in only a single amplicon in a single sample.

Table 2 for coverage per sample per amplicon). In the 0% variant allele sample (pure MCF-7), we confirmed all 13 positions with an alternative allele against rCRS [38] at > 95% allele frequency. At 5/28 positions known to be different between the two cell lines, heteroplasmic variants were observed in all mixture samples (**Supplementary Table 3**), prompting us to omit these positions in further analysis for limit of detection. Thus, we explored 23 positions by SMRT sequencing and confirmed all variant alleles, with a detection limit of 0.1% for 21 positions and a detection limit of 1% for 2 positions (**Table 1** and **Supplementary Figure 6A**). The UltraSEEK method employs amplification of the region(s) of interest by PCR and subsequent detection of the variant(s)-of-interest via a single-base extension using chain terminators labelled with a moiety for solid phase capture, allowing enrichment of product, and identification of the product using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry [36]. By UltraSEEK, we explored 7 positions and detected all variant alleles at those positions, with a detection limit of 0.1% for 5 positions and a detection limit of 1% for 2 positions (**Table 1** and **Supplementary Figure 6B**). In digital PCR, a sample is partitioned into many individual parallel probe-based PCR reactions, each reaction contains either one target molecule or none, allowing a "yes" or "no" answer for the target molecule containing the mutant and wildtype allele in each reaction. By digital PCR 2 positions were evaluated for the variant allele, and one variant allele was detected ≥ 0.01% allele frequency and the other ≥ 0.1% allele frequency (**Table 1** and **Supplementary Figure 6C**).

Detection of de novo mtDNA variants by SMRT sequencing

Since by SMRT sequencing the entire mtDNA was sequenced, we explored all alternative alleles that were called in the dataset of the six sample mixtures containing 0%, 0.001%, 0.01%, 0.1%, 1% and 10% mutant variant frequency. A total of 132 variants were called at 126 positions (some positions contained more than one alternative allele, **Supplementary Table 3**). Besides the documented homoplasmic variants for these two cell lines (35 variants, including the 28 differing alleles described above and 7 concordant alleles), 97 *de novo* variants were detected. Of those, 55 appeared as false positive calls in Integrative Genomics Viewer [39] since they were associated with homopolymer regions or were in close proximity to homoplasmic alternative variants (**Supplementary Figure 5**). Of the remaining 42 *de novo* variants, the allele frequency ranged from 0.01% to 24.8% (**Table 2**). To evaluate if those *de novo* variants are true positive variants or potential false positives, we assessed their validation within the dataset: independent observations of a variant in multiple mixtures, or independent observations of a variant in overlapping regions of the sequenced amplicons. Of the 42 *de novo* variants, 20 were present in multiple mixtures, whereas 22 were present in one mixture only (**Table 2**). Also, 5 had

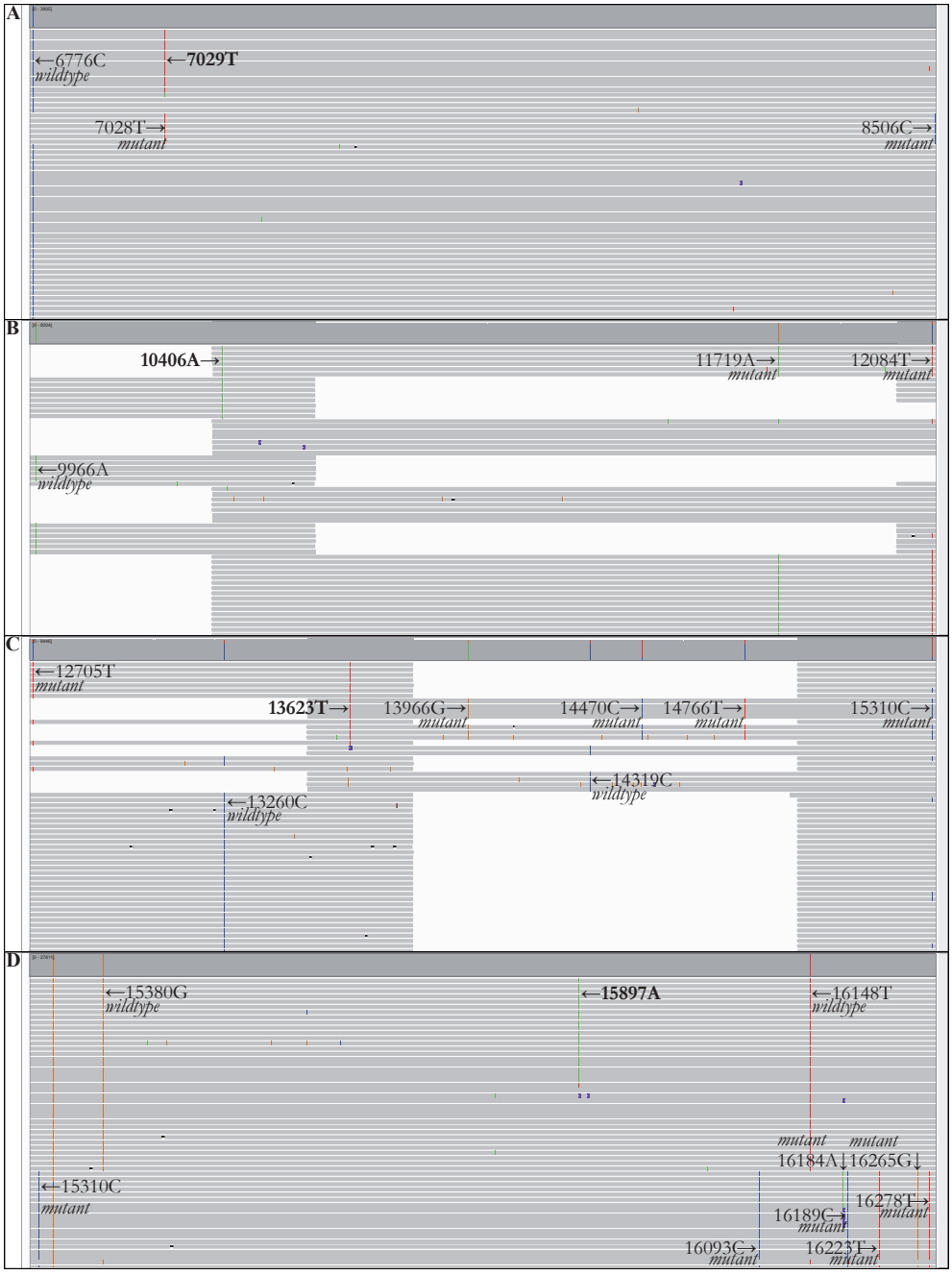


Figure 2 Phasing of *de novo* variants with variants known to belong to either the wildtype (MCF-7) or mutant (MDA-MB-231) genotype, exemplified by four Integrative Genomics Viewer (IGV) screenshots.
A: In the 0.1% mutant sample, position 7029 (T, red) phases together with reads containing the wildtype (MCF-7) variant at position 6776 (C, blue) but not the mutant (MDA-MB-231) variants at positions 7028 (T, red) and 8506 (C, blue). **B**: In the 10% mutant sample, position 10406 (A, green) phases together with reads containing the mutant (MDA-MB-231) variants at position 11719 (A, green) and 12084 (T, red) but not the wildtype (MCF-7) variant at

been detected in the mutant-only sample (100% MDA-MB-231) that was sequenced at lower depth by both SMRT and SBS sequencing (see **Supplementary File**). Ten *de novo* variants were detected in overlapping regions of the sequenced amplicons, and thus represent two independent observations within one sample (**Table 2**). This resulted in 26 *de novo* variants that could be validated in our dataset, and thus true positive calls. A total of 16 *de novo* variants were detected in only a single amplicon in a single sample (**Table 2**), and can in theory be false positive calls (i.e. PCR errors or sequencing errors). These potential false positive variants had an allele frequency between 0.03% and 0.34%. Based on this, if validation of variants in either multiple samples or multiple amplicons is not possible, a conservative threshold on allele frequency for *de novo* variant detection of the SMRT sequencing approach would be ≥ 1.0% allele frequency.

Allelic phasing of mtDNA variants detected by SMRT sequencing

The long read length of SMRT sequencing enables to phase variants i.e. determine if they are present on the same read or on separate reads and thus if they originated from the same or another mtDNA molecule (**Figure 2**). By this, we could evaluate if variants phased together with the known homoplasmic variants of the wildtype (MCF-7) or of the mutant (MDA-MB-231) genotype. Of the 42 *de novo* variants, a total of 32 variants phased together with the wildtype genotype and not with the mutant genotype (**Table 2**). The variants with an allele frequency ≥ 0.5% in the wildtype-only mixture (0% mutant) were typically detected in all mixtures, whereas variants ≤ 0.5% allele frequency in the wildtype-only mixture were typically detected in the mixtures with only low mutant fractions (**Table 2**), hence the detection limit of the method. The remaining 10 *de novo* variants phased together with the mutant genotype and not with the wildtype genotype. Among those 10 variants that phased together with the mutant genotype, were the five that had also been detected in the mutant-only sample (100% MDA-MB-231) sequenced at lower depth by both SMRT and SBS sequencing (see **Supplementary File**). Also here, variants with a higher allele frequency in the mutant-only sample were typically detected

position 9966 (A, green). Note that position 10406 is covered by two amplicons, and thus detected by two independent observations. **C**: In the 10% mutant sample, position 13623 (T, red) phases together with reads containing the mutant (MDA-MB-231) variants at position 12705 (T, red), 13966 (G, orange), 14470 (C, blue), 14766 (T, red) and 15310 (C, blue) but not the wildtype (MCF-7) variants at position 13260 (C, blue) and 14319 (C, blue). Note that position 13623 is covered by two amplicons, and thus detected by two independent observations. **D**: In the 0.1% mutant sample, position 15897 (A, green) phases together with reads containing the wildtype (MCF-7) variants at position 15380 (G, orange) and 16148 (T, red) but not the mutant (MDA-MB-231) variants at position 15310 (C, blue), 16093 (C, blue), 16184 (A, green), 16189 (C, blue), 16223 (T, red), 16265 (G, orange) and 16278 (T, red). Horizontal is the DNA sequence, vertical the individual reads, and alignments sorted by base. Note that the position in IGV does not correspond to the rCRS position due to the use of an extended reference for alignment (see Materials and Methods). INDELs < 2 bases are hidden for clarity.

in the mixtures with high mutant fractions (**Table 2**), hence the detection limit of the method. Thus, by SMRT sequencing we were able to evaluate the origin of the 42 *de novo* variants, phased to either the wildtype or mutant genotype (**Table 2**).

Discussion

In this research, we aimed to develop a sensitive procedure to detect low-frequent single-nucleotide mtDNA variants from frozen tumour tissue. In assessing tumour cell characteristics, tissue specimens are an important source to detect tumour-specific variants. Especially when the focus is on low-frequent variants, frozen tissue is more suitable than formalin-fixed paraffin-embedded tissue since the latter is prone to deamination artefacts [40]. We started by establishing an extraction procedure to obtain mtDNA as pure as possible from frozen tumour tissue. The optimal method was DNA from cytosol fractions (method IV) treated with exonuclease, and resulted in a 270-fold mtDNA enrichment when compared to total cellular DNA extraction (27% versus 0.1% mtDNA yield, **Figure 1**). The method based on alkaline extraction that is normally applied to extract plasmid DNA has also been described by others for preparation of mtDNA-enriched samples [28, 30, 32, 33]. In line with the work by Quispe-Tintaya et al [33], we find for frozen cultured cells a good mtDNA enrichment compared to total cellular DNA extraction (158-fold, **Supplementary Figure 1**). However, application to frozen tumour tissue resulted in only a 5-fold mtDNA enrichment (**Figure 1**) indicating that this method is less suited for frozen specimens. The method that extracts DNA from isolated mitochondria has also been described by others [34], for which we find for frozen cultured cells similar mtDNA enrichment levels compared to total cellular DNA extraction (3-fold, **Supplementary Figure 1**). However, again for frozen tumour tissue we observe lower mtDNA enrichment (2-fold, **Figure 1**). Note that, although the alkaline-based and mitochondria-based extraction methods were equivalent, different methods were applied to extract total cellular DNA in the above mentioned studies, and even among silica-based extraction methods mtDNA yield can be different [41, 42]. Importantly, DNA from cytosol fractions either with or without exonuclease treatment compared to total cellular DNA extraction did also show better results for cultured cells (resp. 33-fold and 760-fold enrichment, **Supplementary Figure 1**). Thus, generally, extraction methods that significantly enrich for mtDNA from frozen cultured cells (and possibly also blood cells) do not guarantee a proper enrichment for mtDNA from frozen tissue.

A high fraction of mtDNA obtained within the DNA extract is vital to minimize the presence of NUMTs, which may lead to misinterpretation of mtDNA variants. Due to the variable number of mtDNA molecules per cell and the variable frequency of

NUMTs, estimating the potential misinterpretation with NUMTs is difficult and unique for each position in each individual. Since the generation of NUMTs is an ongoing process [17-19] estimating NUMT frequency is even more difficult for tumour cells since, they contain all private and all somatic NUMT events that have occurred during tumorigenesis and before that time. This is why we have chosen – and recommend – to analyse a mtDNA extract as pure as possible in SMRT sequencing. Exemplifying, in the case of 20x abundance of a NUMT (which is the case for numerous mtDNA regions [24]) in a cell type with 500 mtDNA molecules, it is possible to misinterpret the NUMT as a mtDNA variant with 8% heteroplasmy (2 x 20 / 500) in a total cellular DNA extract. Indeed, misinterpretation of non-identical mtDNA and NUMT positions is not a rare event and multiple examples have been highlighted in the literature [16, 20-23]. Therefore, obtaining a high mtDNA fraction corresponds to obtaining a high number of mtDNA molecules as opposed to nDNA molecules, decreasing the variant allele frequency of the NUMTs, thus diminishing the likelihood for misinterpretation: a 270-fold increase in mtDNA for the example mentioned above would result in suppressing the NUMT variant to 0.03% heteroplasmy (2 x 20 / 270 x 500).

To detect low-frequent variants in mtDNA, we compared three state-of-the-art approaches. All three methods – SMRT sequencing, UltraSEEK, digital PCR – obtained 100% sensitivity at 1% variant allele frequency (**Table 1**). Specifically, SMRT shows a sensitivity of 100% at 1% allele frequency, 91% at 0.1% allele frequency and 0% at 0.01% allele frequency. SMRT sensitivity mainly depends on the read depth: positions 6221 and 6371 were sequenced less deep and had a detection limit of 1% (**Supplementary Table 3**). UltraSEEK shows a sensitivity of 100% at 1% allele frequency, 71% at 0.1% allele frequency and 0% at 0.01% allele frequency. Digital PCR shows a sensitivity of 100% at 0.1% allele frequency, of 50% at 0.01% allele frequency and 0% at 0.001% allele frequency. Notably, whereas UltraSEEK and digital PCR are limited to the positions chosen beforehand, the SMRT sequencing approach is able to evaluate the entire mtDNA. Since to date no mutational hotspot regions have been described for mtDNA in primary tumour specimens [12-14], this is a valuable feature to study tumour-specific mtDNA variants. A limitation of all three methods is that they start with PCR amplification, and due to the large variation in mtDNA between and within individuals, primer binding sites can encounter variants that can bias PCR amplification. A whole genome sequencing method would enable a more unbiased approach, where a DNA sample is fragmented and subsequently sequenced independent of variants present in the sample. However – as shown by our results using whole-genome sequencing-by-synthesis (SBS) – this method requires deeper sequencing since a substantial part of the reads will be derived from nDNA. A bioinformatics approach would also be needed to filter reads originating from known NUMTs. In addition, the observed spread in mtDNA percentage in DNA extracts from

frozen tumour tissue (**Figure 1**) will lead to variability in the proportion of mtDNA reads between specimens when using a whole genome sequencing approach. This variability is likely due to biological variability in the number of mtDNA molecules within a cell or biochemical differences (e.g. fat or stromal content) between specimens, or due to technical variability in the multiplex qPCR assay. Samples with an extreme high mtDNA:nDNA ratio (and thus those greatly enriched for mtDNA) will have their mtDNA Ct value at the upper end whereas the nDNA Ct will be at the lower end, making the ratio estimation more variable because Ct estimations are less reliable. Also, the observed number of duplicated reads in SBS (87%) is within the expected range for single-end sequencing of the mitochondrial genome. Due to its small size, it contains only 16,569 starting positions for the 776,959 generated reads (**Supplementary Table 1**). When no variants or sequencing errors would be present within the reads, this would result in 97.9% of the reads appearing as duplicate reads. One could also use a targeted approach prior to SBS sequencing. Amplification of the complete mitochondrial genome in a single amplicon has been applied in SBS approaches, obtaining an error rate of 0.33% at a read depth of 20,000x [43]. Sequencing such an amplicon by SMRT is not feasible with the current chemistry, since it would require a read length >80,000 base pairs (5 passes of ~16,569 base pairs). Our targeted approach to amplify mtDNA by primer sets to generate amplicons between 1,700 base pairs and 3,000 base pairs does allow for high quality SMRT reads (≥ 5 passes to create a consensus sequence, minimizing sequencing errors) covering the complete amplicon, and simultaneously minimizes the risk of NUMT amplification (87% of known NUMTs are mtDNA fragments ≤ 1,500 base pairs [22]). In addition, the used primer sets did not generate an amplification product in mtDNA-depleted counterparts of two cell lines (**Supplementary Figure 3**) nor products by *in silico* BLAST, affirming that known NUMTs are unlikely to interfere. A drawback is that template amplification by PCR can introduce errors that may result in false positive calls. To decrease this, the PCR used a high fidelity polymerase (error rate of ~10^{-7}) and the number of PCR cycles was limited (15 + 5 cycli). This would theoretically mean that 98.5-97.5% of the generated products per amplicon are entirely error-free, or that each product contains 0.02 random errors. By setting alternative allelic calls to at least 5 independent high-quality reads we intent to minimize calling PCR errors. An alternative would be to employ molecular barcodes prior to PCR amplification, which will allow tracing PCR duplicates and thus yield more confident calls of the original molecules. Note that five of the *de novo* variants detected by SMRT present in only a single sample appeared on two amplicons and are thus independent observations and unlikely to be PCR errors (**Table 2**). For the *de novo* variants that appear in only one sample on one amplicon (n = 16) we cannot rule out that they are not PCR errors, despite their phasing with a particular genotype (**Table 2**). All those were low-frequent variants (allele frequency between 0.03% and 0.34%). Thus, given

the 100% sensitivity at 1% allele frequency, the SMRT approach is able to call variants reliable ≥ 1% allele frequency. To ascertain that variants below 1% allele frequency are true variants, validation is necessary by either independent re-sequencing (an additional sample, or in some cases in overlapping regions of amplicons within the same sample) or an orthogonal method. Both UltraSEEK and digital PCR prove suitable as orthogonal methods to confirm allelic calls, since they are both able to detect low-frequent variants. Analysis by UltraSEEK can be performed in multiplex (up to hundreds): the region(s) of interest are PCR amplified and subsequently the variant(s)-of-interest are detected via a single-base extension using chain terminators labelled with a moiety for solid phase capture, enrichment of product, and identification using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. However, both UltraSEEK and digital PCR are not suitable for *de novo* variant detection because they do need information on the variants-of-interest beforehand. Also, primer design has to be done for each variant separately, which can be limiting due to design constraints. The sensitivity of UltraSEEK mainly depends on the number of molecules analysed, where 3 variant copies would suffice for detection (corresponding to at least 3,000 total copies for a 0.1% allele frequency). Analysis by digital PCR can be performed in multiplex (up to 4-8), with for each DNA molecule the region of interest is PCR-amplified and subsequently detected by specific probes on the variant-of-interest. Also in here, sensitivity mainly depends on the number of input molecules (minimal 2 variant copies of ≤ 20,000 total copies). The SMRT sequencing approach is as performant in terms of sensitivity (dependent on minimal 5 alternative reads) compared to these two methods, but is not limited to the necessity of knowing positions of variants-of-interest beforehand.

To conclude, our sensitive procedure to detect low-frequent single-nucleotide mtDNA variants from frozen tumour tissue is based on the extraction of DNA from cytosol fractions followed by exonuclease treatment to obtain high mtDNA yield, and subsequent SMRT sequencing for (*de novo*) detection and allelic phasing of variants. Orthogonal validation of variants can be done by either UltraSEEK (in the case of numerous variants) or digital PCR (in the case of a few variants). We conclude that the presented approach enables mtDNA-specific detection of *de novo* variants ≥ 1 % allele frequency.

Materials and methods

Specimens

Cell lines MDA-MB-231 and MCF-7 were cultured using RPMI (*Invitrogen*) supplemented with FBS (10%) (*Lonza*), 100 U/mL penicillin (*Invitrogen*), 100 µg/mL streptomycin (*Invitrogen*) and 0.05 mg/mL gentamycin (*Invitrogen*). A mtDNA-depleted

MDA-MB-231 breast cancer cell line (MDA-MB-231-ρ0) was established by culturing MDA-MB-231 cells in the presence of 50 ng/μL ethidium bromide for 100 days in medium supplemented with uridine (0.05 mg/mL) (*Sigma-Aldrich*) and pyruvate (1 mM) (*Invitrogen*). Frozen 143B and 143B-ρ0 osteosarcoma cell line pellets were kindly provided by dr. W.N.M. Dinjens (Department of Pathology, Erasmus MC). Fresh frozen primary breast tumour tissue specimens (resection material) were selected from the tumour biobank at the Erasmus MC (n = 10, stored in liquid nitrogen). The use of these patient materials was approved by the medical ethics committee of the Erasmus MC (MEC 02.953) and in accordance to the code of conduct of Federation of Medical Scientific Societies in the Netherlands. In the Netherlands, according to the Code of Conduct, informed consent is not required for retrospective analysis of bio-specimens retrieved during standard of care procedures.

DNA extraction and mtDNA enrichment

Input for frozen tumour tissue was standardized at 20 cryosections of 30 μm thickness, which resulted in an average input of 19.2 mg (range of 5.9 - 33.4 mg) tumour tissue per extraction. Input for cultured cells was standardized at 1 million frozen cells per extraction. Total cellular DNA was extracted using the NucleoSpin Tissue kit (*Macherey-Nagel*) according to the supplier's protocol (method I). Alkaline-based extraction was performed using the QIAprep Spin Miniprep kit (*Qiagen*), according to the supplier's protocol (method II). Mitochondria were extracted using the Qproteome mitochondria isolation kit (*Qiagen*) according to the supplier's protocol, and subsequently DNA was extracted using the NucleoSpin Tissue kit (above) (method III). To remove cell nuclei, samples were lysed using detergent that dissolves the cellular membrane (1 mL of 0.5x TBE containing 0.5% (v/v) Triton X-100 [37]) for 10 minutes, followed by sedimentation of the nuclei at 1020 x g for 10 minutes. From the remaining supernatant – the cytosol fraction – DNA was extracted using the QIAamp Circulating Nucleic Acid Kit (*Qiagen*) according to the suppliers' protocol (method IV). In experiments to remove linear DNA, extracts (max. 100 ng DNA) were treated with 40 units of the ATP-dependent exonuclease PlasmidSafe (*Epicentre*) for 3 hours at 37°C. Exonuclease was heat-inactivated (30 minutes 70°C) and the circular DNA was purified using ethanol precipitation (70% ethanol).

DNA quantification and mtDNA purity assessment

All DNA extracts were quantified using the Qubit dsDNA HS assay kit (*Life Technologies*) according to the suppliers' protocol. Purity of mtDNA was assessed in duplicate runs of a multiplex qPCR assay targeting a nuclear and a mitochondrial encoded gene to calculate the ratio of mtDNA molecules opposed to nDNA molecules by the relative quantitation

method ($2^{\Delta Cq}$) as described before [44]. The percentage of mtDNA in the DNA extract was quantified (**Formula 1**) based on the ratio mtDNA:nDNA molecules and the sizes of the mitochondrial reference genome (16,569 base pairs, NC_012920) and complete reference genome (haploid 3,088,269,805 base pairs, GRCh38). If no amplification signal for the nuclear encoded gene was obtained, the ratio mtDNA:nDNA was set to 20,000,000 corresponding to a mtDNA percentage of 99%.

Formula 1

$$mtDNA\ percentage = \frac{ratio * mitochondrial\ genome\ size}{(ratio * mitochondrial\ genome\ size) + nuclear\ genome\ size} * 100$$

Whole genome sequencing-by-synthesis (SBS)

Input DNA was mechanically sheared using focused-ultrasonicator (*Covaris*) to yield fragments of ~ 300 base pairs in length, which required the following shearing-time for different DNA extracts: 90 seconds for total cellular DNA, 120 seconds for total cellular DNA treated with exonuclease, 90 seconds for cytosol fraction DNA, 50 seconds for cytosol fraction DNA treated with exonuclease. Sequence library was created using the Thruplex DNA-seq sample preparation kit (*Rubicon Genomics*), using 0.1-7.7 ng sheared input DNA. Sequencing was performed on an Illumina HiSeq2500 sequencer using HiSeq Rapid v2 chemistry and yielding 100 nucleotides single-end reads.

UltraSEEK

UltraSEEK assays were designed using the AgenaCx online assay design software which automatically selects the PCR and extension primers (**Supplementary Table 4**), and adds to each reaction control assays for PCR and capturing. All oligonucleotides were obtained from Integrated DNA Technologies and control oligo's from Agena Bioscience GmbH. Reactions were performed as described before [36], using reagents obtained from Agena Bioscience. Briefly, PCR (45 cycles) was followed by shrimp alkaline phosphatase treatment and single base primer extension using biotinylated ddNTPs specific for the mutant alleles. After capture of the extended primers using streptavidin-coated magnetic beads, a cation-exchange resin was added for cleaning and 10-15 nl of the reaction was transferred to a SpectroCHIP® Array (a silicon chip with pre-spotted matrix crystals) using an RS1000 Nanodispenser (*Agena Bioscience*). Data were acquired via matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using a MassARRAY Analyser 4 (*Agena Bioscience*). After data processing, a spectrum was produced with relative intensity on the y-axis and mass/charge on the x-axis. Typer Analyser software was used for data analysis and report generation.

Digital PCR

Custom assays for two alternative variants were performed on the Quantstudio 3D digital PCR system (*Thermo Fisher*) according to the supplier's protocol, with an adaption to the DNA input due to high mtDNA copy number. Reactions contained 20 pg of DNA in 1x dPCR mastermix v2, 0.9 µM of each primer (*Invitrogen*) and 0.2 µM of each probe (*Sigma*) (**Supplementary Table 4**). After initial denaturation for 10 minutes at 96°C, the 40-cycle two-step PCR was performed at 30 seconds denaturation (98°C) and 120 seconds annealing/extension (56°C), and followed by a final 2 minute extension (56°C). To calculate a variant frequency of the alternative variant, the threshold for signal dots was set to at least two dots.

Single Molecule Real-Time (SMRT) sequencing

Amplicons covering the complete mtDNA [45, 46] (**Supplementary Table 4**) were generated in singleplex PCR reactions with initial denaturation for 3 minutes at 98°C, 15 cycles of a three-step PCR with 10 seconds denaturation (98°C), 30 seconds annealing (67°C) and 90 seconds extension (72°C), and final extension (72°C) for 5 minutes. Each 50 µL reaction contained 2.5 ng of template DNA and 1 unit of Hot-Start Q5 High Fidelity DNA polymerase (*NEB*) in 1x Q5 reaction buffer, 200 µM dNTPs and 0.5 µM of each 5'-M13 tailed primer (*Invitrogen*) (**Supplementary Table 4**). Specificity of the generated products was confirmed using microchip electrophoresis (DNA-12000 reagent kit, *Shimadzu*). Amplicons were equimolar pooled per sample and purified using AMPure PB paramagnetic beads (*Pacific Biosciences*) with a 0.6 beads:sample ratio according to the SMRTbell Template Prep Kit protocol and eluted in 10 mM Tris-HCl pH 8.5. The 5'-M13 universal sequence tail of the primers allowed barcoding of each sample by performing 5 amplification cycles of the three-step PCR as described above but with an annealing temperature of 58°C. Specificity of the generated products was confirmed using microchip electrophoresis (BioAnalyzer, DNA12000 or High Sensitivity DNA kit, *Agilent*). A final mix of barcoded fragments of all samples was obtained by equimolar pooling and subsequently purified using AMPure PB paramagnetic beads with a 0.6 beads:sample ratio. Concentration of the final mix was determined using the Qubit dsDNA HS assay kit, and SMRTbell library was generated according to the Amplicon Template Preparation and Sequencing guide (*Pacific Biosciences*). Sequencing was performed on Pacific Biosciences RSII with P6-C4 sequencing chemistry and 360 minutes movie-time or Sequel platforms with version 2 sequencing chemistry and 600 minutes movie-time. A total of twenty-two RSII and two Sequel SMRT cells were used to reach a read depth estimated at 3000x per sample. In addition, two RSII SMRT cells were used to reach an estimated 5000x for one sample (cell line mixture with 0.1% mutant allele frequency).

Bioinformatics

Whole genome sequencing-by-synthesis (SBS) reads were trimmed and aligned using hisat2 [47] against the human reference genome GRCh38, after which the percentage of mtDNA was calculated (**Formula 2**). In addition, for evaluation of detected variants (**Supplementary File**), SBS reads were aligned against an extended version of rCRS (BWA-MEM version 0.7.15 default parameters [48]) and duplicate reads marked (Picard MarkDuplicates default parameters http://broadinstitute.github.io/picard/). We aligned the data against extended versions of rCRS (**Supplementary Table 5**) to compensate for mapping bias due to circularity of the mitochondrial genome.

Formula 2

$$percentage\ reads\ of\ mitochondrial\ origin = \frac{aligned\ reads\ on\ chrM}{aligned\ reads\ on\ GRCh38} * 100$$

Single Molecule Real-Time (SMRT) sequencing RS bax.h5 files were converted to Sequel BAM files, of which circular consensus reads (CCS) were generated using the CCS2 algorithm for each sample-specific barcode [49]. Next, a minimum quality threshold of 99% and at least five passes of the SMRTbell were applied to select for highly accurate single-molecule reads. Selected CCS reads were trimmed (Cutadapt [50] for primers-tails) and subsequently aligned against an extended rCRS (BWA- MEM version 0.7.15 parameters -k17 -W40 -r10 -A1 -B1 -O1 -E1 -L0 [48]). We aligned the data against extended versions of rCRS (**Supplementary Table 5**) to compensate for mapping bias due to circularity of the mitochondrial genome.

For the comparison between SBS and SMRT sequencing methods (Supplementary File), pileup files were generated (Bioconductor Rsamtools 1.26.2 pileup function with pileupParam min_base_quality=30, min_mapq=0, min_nucleotide_depth=0, min_minor_allele_depth=0, distinguish_strands=TRUE, distinguish_nucleotides=TRUE, ignore_query_Ns=TRUE, include_deletions=FALSE, include_insertions=FALSE and in the case of SBS data flag isDuplicate=FALSE) and converted back to rCRS positions. For evaluation of detection limit and *de novo* variant detection for SMRT data, pileup files were generated as described above but with a more stringent threshold on the minimal number of alternative allele reads (min_nucleotide_depth=5) to minimize detection of potential PCR errors (see **Supplementary File**). All detected variants were manually inspected in the Integrative Genomics Viewer (IGV, *Broad Institute*) [39]. Phasing of variants was done by manual inspection of every read containing the detected alternative variant and evaluating the other detected alternative variants present on that read.

MDA-MB-231 and MCF-7 mitochondrial sequences were obtained from the NCBI GenBank (resp. AB626609.1 and AB626610.1, deposited after resequencing

by Imanishi et al [38]) and blasted against rCRS to obtain the homoplasmic mtDNA positions alternative to the reference sequence for these two cell lines (NCBI's nucleotide web blast, https://blast.ncbi.nlm.nih.gov).

Data availability

Sequencing datasets can be accessed as BAM files (.bam) from the European Nucleotide Archive under accession number PRJEB23243.

Supplementary data

Supplementary data for this article are available online at Scientific Reports (https://www.nature.com/srep/).

References

1. Robin, E.D. and R. Wong, *Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells.* J Cell Physiol, 1988. **136**(3): p. 507-513.
2. Wiesner, R.J., J.C. Ruegg, and I. Morano, *Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues.* Biochem Biophys Res Commun, 1992. **183**(2): p. 553-559.
3. Legros, F., et al., *Organization and dynamics of human mitochondrial DNA.* J Cell Sci, 2004. **117**(13): p. 2653-2662.
4. Attimonelli, M., et al., *HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research.* BMC Bioinformatics, 2005. **6**(S4): S4.
5. Samuels, D.C., et al., *Recurrent tissue-specific mtDNA mutations are common in humans.* PLoS Genet, 2013. **9**(11): e1003929.
6. He, Y., et al., *Heteroplasmic mitochondrial DNA mutations in normal and tumour cells.* Nature, 2010. **464**(7288): p. 610-614.
7. Li, M.K., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(8): p. 2491-2496.
8. Calloway, C.D., et al., *The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age.* Am J Hum Genet, 2000. **66**(4): p. 1384-1397.
9. Schon, E.A., S. DiMauro, and M. Hirano, *Human mitochondrial DNA: roles of inherited and somatic mutations.* Nature Reviews Genetics, 2012. **13**(12): p. 878-890.
10. Chatterjee, A., E. Mambo, and D. Sidransky, *Mitochondrial DNA mutations in human cancer.* Oncogene, 2006. **25**(34): p. 4663-4674.
11. Wallace, D.C., *Mitochondria and cancer.* Nat Rev Cancer, 2012. **12**(10): p. 685-698.
12. Larman, T.C., et al., *Spectrum of somatic mitochondrial mutations in five cancers.* Proc Natl Acad Sci U S A, 2012. **109**(35): p. 14087-14091.
13. Ju, Y.S., et al., *Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer.* Elife, 2014. **3**: e02935.
14. Stewart, J.B., et al., *Simultaneous DNA and RNA mapping of somatic mitochondrial mutations across diverse human cancers.* PLoS Genet, 2015. **11**(6): e1005333.
15. Blanchard, J.L. and G.W. Schmidt, *Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns.* Mol Biol Evol, 1996. **13**(3): p. 537-548.
16. Hazkani-Covo, E., R.M. Zeller, and W. Martin, *Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes.* PLoS Genet, 2010. **6**(2): e1000834.
17. Caro, P., et al., *Mitochondrial DNA sequences are present inside nuclear DNA in rat tissues and increase with age.* Mitochondrion, 2010. **10**(5): p. 479-486.
18. Ju, Y.S., et al., *Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells.* Genome Research, 2015. **25**(6): p. 814-824.
19. Dayama, G., et al., *The genomic landscape of polymorphic human nuclear mitochondrial insertions.* Nucleic Acids Res, 2014. **42**(20): p. 12640-12649.
20. Parfait, B., et al., *Co-amplification of nuclear pseudogenes and assessment of heteroplasmy of mitochondrial DNA mutations.* Biochem Biophys Res Commun, 1998. **247**(1): p. 57-59.
21. Parr, R.L., et al., *The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation.* BMC Genomics, 2006. **7**: 185.
22. Ramos, A., et al., *Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies.* Mitochondrion, 2011. **11**(6): p. 946-953.
23. Albayrak, L., et al., *The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome.* BMC Genomics, 2016. **17**(1): 1017.

24. Cui, H., et al., *Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders.* Genetics in Medicine, 2013. **15**(5): p. 388-394.
25. Alexandrov, L.B. and M.R. Stratton, *Mutational signatures: the patterns of somatic mutations hidden in cancer genomes.* Curr Opin Genet Dev, 2014. **24**: p. 52-60.
26. Helleday, T., S. Eshtad, and S. Nik-Zainal, *Mechanisms underlying mutational signatures in human cancers.* Nat Rev Genet, 2014. **15**(9): p. 585-598.
27. Palva, T.K. and E.T. Palva, *Rapid isolation of animal mitochondrial DNA by alkaline extraction.* FEBS Lett, 1985. **192**(2): p. 267-270.
28. Defontaine, A., F.M. Lecocq, and J.N. Hallet, *A rapid miniprep method for the preparation of yeast mitochondrial DNA.* Nucleic Acids Res, 1991. **19**(1): 185.
29. Lindberg, G.L., et al., *Recovery of mitochondrial DNA from blood leukocytes using detergent lysis.* Biochem Genet, 1992. **30**(1-2): p. 27-33.
30. Peloquin, J.J., D.M. Bird, and E.G. Platzer, *Rapid miniprep isolation of mitochondrial DNA from metacestodes, and free-living and parasitic nematodes.* J Parasitol, 1993. **79**(6): p. 964-967.
31. Yamada, Y., et al., *Comparison of different methods for extraction of mitochondrial DNA from human pathogenic yeasts.* Jpn J Infect Dis, 2002. **55**(4): p. 122-125.
32. Graffy, E.A. and D.R. Foran, *A simplified method for mitochondrial DNA extraction from head hair shafts.* J Forensic Sci, 2005. **50**(5): p. 1119-1122.
33. Quispe-Tintaya, W., et al., *Fast mitochondrial DNA isolation from mammalian cells for next-generation sequencing.* Biotechniques, 2013. **55**(3): p. 133-136.
34. Gould, M.P., et al., *PCR-Free enrichment of mitochondrial DNA from human blood and cell lines for high quality next-generation DNA sequencing.* PLoS One, 2015. **10**(10): e0139253.
35. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-138.
36. Mosko, M.J., et al., *Ultrasensitive detection of multiplexed somatic mutations using MALDI-TOF mass spectrometry.* J Mol Diagn, 2016. **18**(1): p. 23-31.
37. van Strijp, D., et al., *Complete sequence-based pathway analysis by differential on-chip DNA and RNA extraction from a single cell.* Sci Rep, 2017. **7**(1): 11030.
38. Imanishi, H., et al., *Mitochondrial DNA mutations regulate metastasis of human breast cancer cells.* PLoS One, 2011. **6**(8): e23401.
39. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-26.
40. Weerts, M.J.A., et al., *Somatic tumor mutations detected by targeted next generation sequencing in minute amounts of serum-derived cell-free DNA.* Sci Rep, 2017. **7**(1): 2136.
41. Guo, W., et al., *DNA extraction procedures meaningfully influence qPCR-based mtDNA copy number determination.* Mitochondrion, 2009. **9**(4): p. 261-265.
42. Andreu, A.L., et al., *Quantification of mitochondrial DNA copy number: pre-analytical factors.* Mitochondrion, 2009. **9**(4): p. 242-246.
43. Zhang, W., H. Cui, and L.J. Wong, *Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing.* Clin Chem, 2012. **58**(9): p. 1322-1331.
44. Weerts, M.J.A., et al., *Mitochondrial DNA content in breast cancer: Impact on in vitro and in vivo phenotype and patient prognosis.* Oncotarget, 2016. **7**: p. 29166-29176.
45. Ramos, A., et al., *Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification.* Electrophoresis, 2009. **30**(9): p. 1587-1593.
46. Ramos, A., et al., *Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the New Human Genome Reference Sequence (GRCh37).* Electrophoresis, 2011. **32**(6-7): p. 782-783.
47. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nat Methods, 2015. **12**(4): p. 357-360.

48. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-595.
49. Anvar, S.Y., et al., *TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes.* Bioinformatics, 2014. **30**(12): p. 1651-1659.
50. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. 10-12.

CHAPTER 7


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Asp
Ser
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Leu

Tumour-specific mitochondrial DNA variants are rarely detected in cell-free DNA

Marjolein J.A. Weerts | Eveline C. Timmermans | Anja van de Stolpe | Rolf H.A.M. Vossen | Seyed Y. Anvar | John A. Foekens | Stefan Sleijfer | John W.M. Martens

Neoplasia 2018; 20(7):687-696

Abstract

The use of blood-circulating cell-free DNA (cfDNA) as a 'liquid-biopsy' in oncology is being explored for its potential as a cancer biomarker. Mitochondria contain their own circular genomic entity (mitochondrial DNA, mtDNA), up to even thousands of copies per cell. The mutation rate of mtDNA is several orders of magnitude higher than that of the nuclear DNA. Tumour-specific variants have been identified in tumours along the entire mtDNA, and their number varies among and within tumours. The high mtDNA copy number per cell as well as the high mtDNA mutation rate makes it worthwhile to explore the potential of tumour-specific cf-mtDNA variants as cancer marker in the blood of cancer patients. We used single-molecule real-time (SMRT) sequencing to profile the entire mtDNA of nineteen tissue specimens (primary tumour and/or metastatic sites, and tumour-adjacent normal tissue) and nine cfDNA samples, originating from eight cancer patients (5 breast, 3 colon). For each patient, tumour-specific mtDNA variants were detected and traced in cfDNA by SMRT sequencing and/or digital PCR to explore their feasibility as cancer biomarker. As a reference, we measured other blood-circulating biomarkers for these patients, including driver mutations in nuclear-encoded cfDNA and cancer-antigen levels or circulating tumour cells. Four of the twenty-four (17%) tumour-specific mtDNA variants were detected in cfDNA, however at much lower allele frequencies compared to mutations in nuclear-encoded driver genes in the same samples. Also, extensive heterogeneity was observed among the heteroplasmic mtDNA variants present in an individual. We conclude that there is limited value in tracing tumour-specific mtDNA variants in blood-circulating cfDNA with the current methods available.

Introduction

Mitochondria are organelles within our cells responsible for a variety of functions, including energy production and initiating apoptosis. Their small circular genome (mitochondrial DNA, mtDNA) encodes for proteins essential in the oxidative phosphorylation system and the transfer RNA and ribosomal RNA molecules of the mitochondrial translation apparatus. Within a single cell, multiple copies of mtDNA exist (mtDNA content), but due to its small size the mtDNA represents only a minor fraction of the total cellular DNA (< 0.1%). In general, cells with high energy demand (e.g. muscle cells) have a higher mtDNA content than cells with lower energy demand (e.g. blood cells) [1]. In human cancer, changes in mtDNA content have been reported when tumour specimens are compared to their normal counterparts [2]. The polyploid nature of mtDNA invokes the concept of only a single (homoplasmy) or two or more mitochondrial genotypes (heteroplasmy) within a cell. It has been shown that heteroplasmy patterns within an individual can differ between tissues, even in an allele-specific manner [3-6]. Also within cancer, tumours harbour mtDNA that is genetically different to their normal counterparts, either at a homo- or heteroplasmic level, and their number and position vary among tumours [7-9]. Interestingly, since the mutation rate of mtDNA is several orders of magnitude higher than that of nuclear DNA (nDNA) [10], it is very informative to assess phylogenetic distance not only intra- and inter-species, but also inter-individual.

Within oncology, the use of blood-circulating cell-free DNA (cfDNA) as a 'liquid-biopsy' is being explored for its potential as a screening tool, to establish prognosis, or as a marker for response to treatment. The origin of cfDNA is mainly from apoptotic cells, hence its typical fragmentation pattern representing DNA cleavage between nucleosomes or chromatosomes (~146-166 base pairs and multiples thereof) [11]. The physical characteristics of cf-mtDNA have not been studied as extensively as its nuclear counterpart. Since mtDNA is packed into nucleoids [12], which are not fragmented during apoptosis [13], the fragmentation pattern as seen for nDNA does not apply to mtDNA. Indeed, the majority of the cf-mtDNA in human plasma appears associated with particles of at least 0.45 μm in diameter [14] and a fraction of it is severely fragmented down to at least 30 base pairs [15-18]. If not fragmented, the circular nature of mtDNA might render it less susceptible to enzymatic cleavage and thus more stable within the circulation. The total amount of cfDNA is often increased in cancer patients compared to healthy individuals, for both DNA from the nucleus as well as mtDNA [16, 19-23]. The detection of tumour-specific cfDNA is aided by the aberrations present in the cancer's genome, and thus by the detection of tumour-specific mutations within the cfDNA. A few studies have attempted to detect mtDNA variants in blood-derived cfDNA [4, 24-29] or other bodily fluids [24, 30-33]. However, in the studies on blood-derived cf-mtDNA,

used methods were either not very sensitive (i.e. conventional Sanger sequencing), or the variants were not truly tumour-derived (i.e. already present in matched normal specimens). Also, quantitative variant allele frequencies were not reported in all assessed samples, making interpretation of these results difficult. Nevertheless, the combination of a high copy number per cell, a high mutation rate and potentially high stability within the circulation make it worthwhile to explore the potential of tumour-specific variants in cf-mtDNA as a cancer biomarker.

In this study, we used a targeted Single-Molecule Real-Time (SMRT) sequencing approach to profile the entire mtDNA of the primary tumour and/or metastatic sites, tumour-adjacent normal tissue and cfDNA of eight cancer patients. We have recently shown that the SMRT sequencing approach is able to reliably detect unknown variants ≥ 1.0% allele frequency and to trace known low-frequent variants down to at least 0.1% allele frequency [34]. In our cohort we observed tumour-specific mtDNA variants for each patient, and explored the feasibility to trace these tumour-specific variants in cf-mtDNA as a cancer biomarker.

Results

We sequenced the entire mtDNA of nineteen tissue specimens and nine cfDNA samples originating from eight cancer patients, including at least one tumour and one cfDNA sample for each patient. For four tissue specimens (primary tumour of P1, P2 and P3, and normal tissue of P1), we performed independent re-sequencing of another (nearby) section of the specimen. For one patient (P1), a total of five consecutive cfDNA samples were profiled for mtDNA variants by dPCR. The study cohort consisted of two cancer types – breast and colon cancer – at variable disease stages. Patient characteristics are summarized in **Table 1**. After haplotyping the mtDNA of each specimen for each individual to make sure that samples were matched correctly (**Supplementary Figure 1**), the heteroplasmic mtDNA variants present in each sample were evaluated.

Heterogeneity in heteroplasmic mtDNA variants between tumour and normal tissue

In the tumour and normal tissue specimens, the number of heteroplasmic mtDNA variants ranged from 0 up to 14 per specimen (**Table 1**), with allele frequencies between 0.2% and 99.4% (**Figure 1**) (**Supplementary Table 3**). Good concordance was observed in detected variants between different sections of a specimen (P1, P2 and P3, **Figure 1**), with two of the eighteen variants missed due to coverage and thus limit of detection (6255G>A at 0.8% allele frequency in P1 normal mammary tissue and 9058A>G at 0.2% allele frequency in P2 primary tumour tissue **Supplementary Table 3**). Heterogeneity was

Table 1 Patient characteristics and number of detected heteroplasmic mtDNA variants for each specimen by SMRT sequencing.

Patient	Primary tumour type	Sex	Age at diagnosis	Stage	Clinicopathological	Specimen	Heteroplasmic mtDNA variants
P1	Breast	Female	75	T4 cN+ → ypN0 Mx	ER: positive PR: positive HER2: not done Primary tumour size: 2 cm	Primary tumour Normal mammary Serum 1 Serum 2 Serum 3 Serum 4 Serum 5	4 3 5 *not done* *not done* *not done* 3
P2	Breast	Female	81	T3 pN2 M0	ER/PR/HER2: not done Primary tumour size: 8 cm	Primary tumour Normal mammary Serum	7 14 4
P3	Breast	Female	57	T3 cN+ → ypN0 M0	ER: positive PR: positive HER2: balanced Primary tumour size: 8.5 cm	Primary tumour Normal mammary Serum	3 4 0
P4	Breast	Female	49	T3 pN1 M0	ER: positive PR: positive HER2: not done Primary tumour size: 5.2 cm	Primary tumour Normal mammary Serum	3 4 2
P5	Breast	Female	64	ND	ER/PR/HER2: unknown	Primary tumour Serum	1 7
P6	Colon	Male	68	T3 pN1 M1	Right hemicolon Dukes: D Differentiation: moderate Primary tumour size: 3.3 cm	Primary tumour Liver metastasis 1 Liver metastasis 2 Normal colon Normal liver Plasma	5 4 7 6 7 5
P7	Colon	Female	63	T4 pN1 M1	Sigmoid Dukes: D Differentiation: moderate Primary tumour size: 5.5 cm	Primary tumour Omental metastasis Normal colon Plasma	1 4 0 1
P8	Colon	Male	84	T3 and T3 pN0 and pN1 Mx	Sigmoid and distal colon Dukes: B and C Differentiation: moderate Primary tumour size: 6 and 5 cm	Liver metastasis Normal liver Plasma	3 5 9

observed between the tumour and normal tissues: majority of the variants (80%) were present either only within the tumour tissue (n = 24) or only within the normal tissue (n = 27) (**Figure 1**). Generally, tumour-only variants had higher allele frequency than

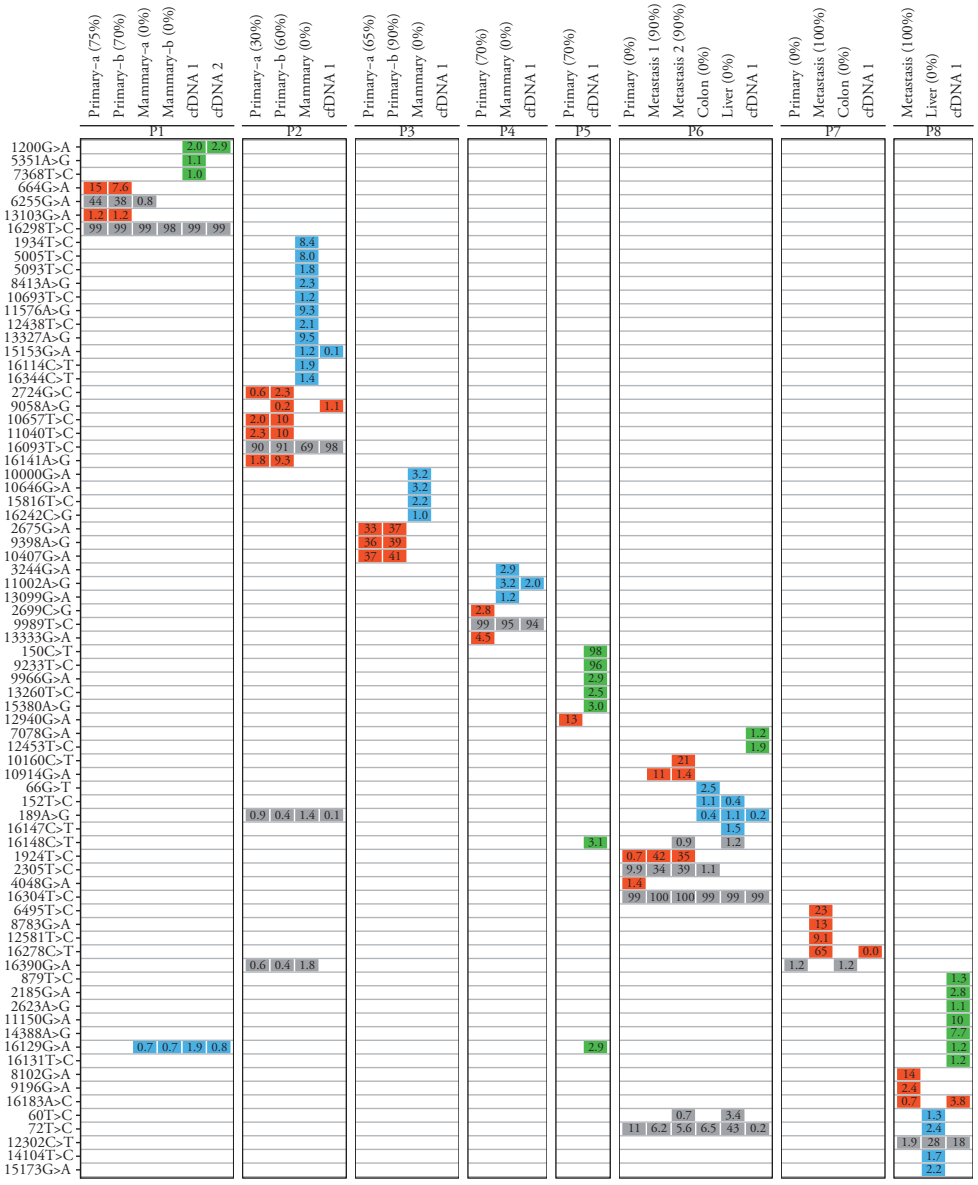


Figure 1 Heteroplasmic mtDNA variants in tumour and normal tissue.
Heteroplasmic mtDNA variants (vertical) detected by SMRT sequencing of tumour and normal tissue (horizontal) of eight cancer patients (P1 to P8). Ubiquitous, normal-only and tumour-only variants in respectively grey, blue and red, cfDNA variants of unknown tissue origin in green. Within the squares allele frequency (%) of the variant is indicated. The percentage of tumour cells in the analysed sections based on morphological estimations in HE-stained slides between brackets behind tissues.

normal-only variants (respectively median (interquartile range IQR) of 9.7% (23.7%) versus 1.9% (1.6%), Mann-Whitney U test P < 0.001).

Also, two of the ubiquitous variants (n = 13) showed heteroplasmic expansion between the normal and the tumour tissue (variants 6255G>A and 2305T>C in respectively P1 and P6, **Figure 1**). Note that some of the ubiquitous variants at low heteroplasmy might not be present in the tumour cells but in normal cells residing the tumour tissue (i.e. variants 189A>G and 16390G>A in P2, 60T>C and 66G>T in P6, and 12302C>T in P8). A phylogenetic relationship based on the tumour-specific mtDNA variants was evident between the primary and the two metastatic tumour sites of P6 (**Figure 2**).

Due to the long read-length of our sequencing approach – in the order of two thousand nucleotides – the detected variants could be grouped based on their presence on the same read or on separate reads and thus we could decipher if they originated from the same or another mtDNA molecule (phasing of variants). A total of 65 combinations of variants were close enough for phasing (**Supplementary Table 4**), of which 19 phased (partly) together and 46 were mutually exclusive. Interestingly, in tumour tissue the heteroplasmic variants 10657T>C and 11040T>C in P2 and variants 9398A>G and 10407G>A in P3 phased together, but variants 1924T>C and 2305T>C in P6 were mutually exclusive.

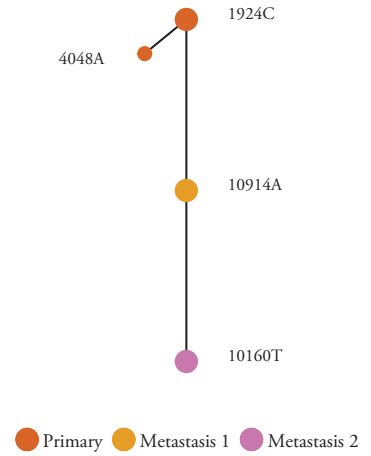


Figure 2 Schematic of phylogenetic relationship between the sequenced colorectal cancer specimens of P6.

Heteroplasmic mtDNA variants within cfDNA

In the cfDNA samples, the number of heteroplasmic mtDNA variants (cf-mtDNA) ranged from 0 up to 9 per sample (**Table 1**), with detected allele frequencies between 0.04% and 99.4% (**Figure 1**) (**Supplementary Table 3**). Majority of the detected cf-mtDNA variants (59%) were not detected in the corresponding tissues we evaluated, and thus are of unknown tissue origin (n = 20). Some of the variants were detected within only

the normal tissue (n = 5) or both the tumour and normal tissue (n = 8), indicating that these are heteroplasmic patient-specific but not tumour-specific heteroplasmic mtDNA variants present as cf-mtDNA in the circulation. Of the twenty-four tumour-specific and two tumour-expanded heteroplasmic mtDNA variants present in the tumour specimens, only three were detected by sequencing the cfDNA (**Figure 1**): in P2 variant 9058A>G was present at 0.2% allele frequency in one of the two replicates of the primary tumour and 1.1% allele frequency in the cfDNA, in P7 variant 16278C>T was present at 64.7% allele frequency in the liver metastasis and 0.04% allele frequency in the cfDNA, and in P8 variant 16183A>C was present at 0.7% allele frequency in the liver metastasis and 3.8% allele frequency in the cfDNA (**Supplementary Table 3**). Note that in P2 and P8 the heteroplasmy level of the variant in the tumour tissue was very low. We confirmed by dPCR (orthogonal technique) the absence of one high-frequent tumour-specific or one high-frequent tumour-expanded variant in the sequenced cfDNA samples for both P1 and for P6 (**Table 2**). Note that the variant allele frequency of those variants in the tissue samples was comparable between SMRT sequencing and dPCR detection. For P1, we extended the number of cfDNA samples by three sera at different time-points. The cancer-antigen level in these three sera was extremely high (**Figure 1**, **Table 3**) indicative for a high tumour load at that point in time. In this patient, we detected by dPCR at low variant allele frequency the tumour-expanded cf-mtDNA variant prior to start of hormonal therapy (6255G>A 0.03% allele frequency) and both the tumour-expanded

Table 2 Heteroplasmic mtDNA variants detected by dPCR in two patients.

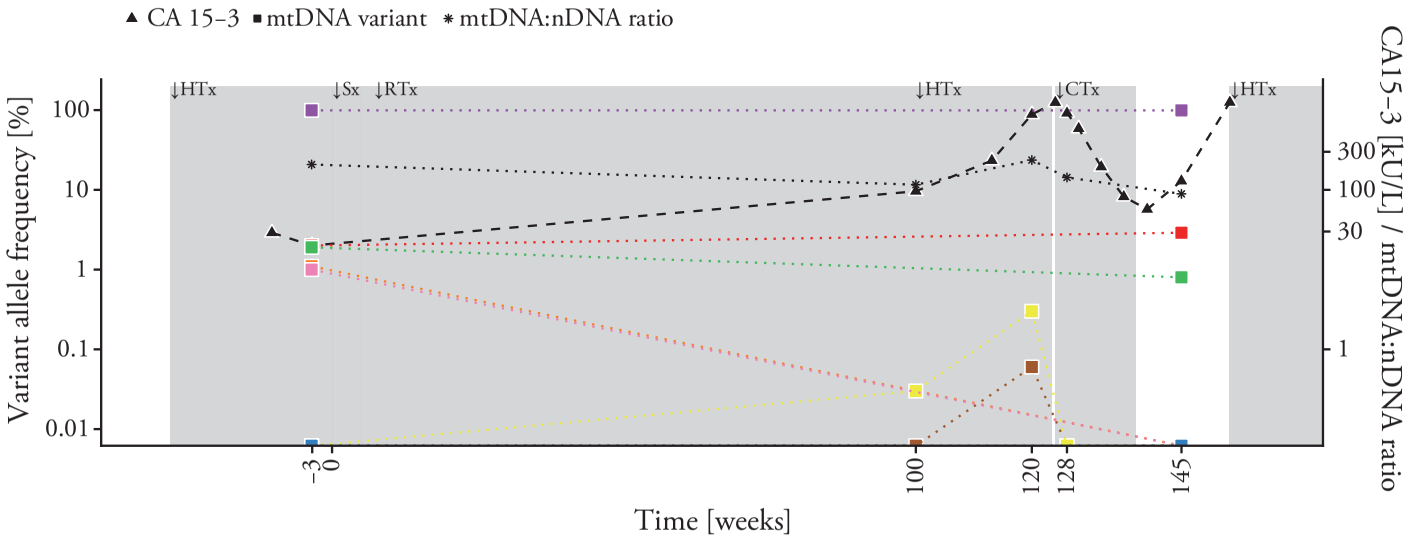
Patient	Tumour type	Specimen	Variant 664G>A (allele frequency)	Variant 6255G>A (allele frequency)
P1	Breast	Primary tumour-a	12.9%	46.2%
		Primary tumour-b	5.4%	35.9%
		Normal mammary-a	0.01%	0.9%
		Normal mammary-b	0.05%	0.5%
		Serum 1	*nd*	*nd*
		Serum 2*	*nd*	*0.03%*
		Serum 3*	*0.06%*	*0.3%*
		Serum 4*	*nd*	*nd*
		Serum 5	*nd*	*nd*
Patient	Tumour type	Specimen	Variant 1924T>C (allele frequency)	Variant 2305T>C (allele frequency)
P6	Colon	Primary tumour	0.9%	10.4%
		Liver metastasis 1	36.2%	40.1%
		Liver metastasis 2	46.2%	34.2%
		Normal colon	0.3%	1.1%
		Normal liver	0.08%	0.05%
		Plasma	*nd*	*nd*

Asterisks indicate the samples that had not been analysed by SMRT sequencing. nd indicates not detected.

Table 3 Blood-based markers in sera or plasma of the eight evaluated cancer patients.

Patient	Blood draw	Tumour sites in situ at blood draw	Cancer antigen		Circulating tumour cells	Tumour cf-nDNA [allele frequency]	Tumour cf-mtDNA [allele frequency]
			CA15.3	CA125			
P1 (breast)	1	Primary	20 kU/L	*not done*	*not done*	*not done*	0%
	2	Metastases (bone, lung, liver)	96 kU/L	*not done*	*not done*	*not done*	0.03%
	3	Metastases (bone, lung, liver)	883 kU/L	*not done*	*not done*	0%[a]	0.06 - 0.3%
	4	Metastases (bone, lung, liver)	587 kU/L	*not done*	*not done*	*not done*	0%
	5	Metastases (bone, lung, liver)	129 kU/L	*not done*	*not done*	*not done*	0%
P2 (breast)	1	Metastases (bone, lung)	30 kU/L	*not done*	*not done*	0.06%	1.1% (?)
P3 (breast)	1	Metastases (colon, spleen, pancreas, omentum)	*not done*	97 kU/L	*not done*	0.3%[a]	0%
P4 (breast)	1	Primary and metastasis (lymph nodes)	*not done*	9 kU/L	*not done*	0%[a]	0%
P5 (breast)	1	Metastasis (bone)	34 kU/L	*not done*	*not done*	47.5%	0%
P6 (colon)	1	Primary and metastases (lymph nodes, liver)	*not done*	*not done*	2 / 7.5 mL	2.4%	0%
P7 (colon)	1	Primary and metastases (lymph nodes, liver, small intestine, omentum)	*not done*	*not done*	0 / 7.5 mL	13.4 - 18.5%	0.04%
P8 (colon)	1	Metastasis (liver)	*not done*	*not done*	35 / 7.5 mL	7.8 - 15.0%	3.8% (?)

[a]Note that for these samples not the entire nDNA was evaluated, but only the subset of driver genes covered by the Oncomine Breast cfDNA Assay.



ratio [mtDNA:nDNA]	208		116	236 143	89
CA15–3 [kU/L]	20		96	883 587	129
7368C [%]	1		ND	ND ND	0
664A [%]	0		0	0.06 0	0
6255A [%]	0		0.03	0.3 0	0
5351G [%]	1.1		ND	ND ND	0
16298C [%]	99.2		ND	ND ND	99.1
16129A [%]	1.9		ND	ND ND	0.8
13103A [%]	0		ND	ND ND	0
1200A [%]	2		ND	ND ND	2.9

Figure 3 Timeline of patient 1.
The allele frequency of detected mtDNA variants (squares), levels of CA 15-3 (triangles), and ratio of mtDNA:nDNA molecules (asterisks) on log-scale in the cfDNA samples (vertical) at five time points (horizontal). The table provides the specifics per variable (color-coded). At the top of the graph treatment is indicated (Sx: surgery, RTx: radiotherapy, HTx: hormone therapy, CTx: chemotherapy), where time = 0 corresponds to the surgery of the primary tumour. A grey background indicates that the patient was receiving systemic therapy. Note that allele frequency in the first and last sera were evaluated by SMRT sequencing, whereas in the second, third and fourth sera by dPCR. ND not done.

and tumour-specific cf-mtDNA variants prior to start of chemotherapy (6255G>A 0.3%, 664G>A 0.06% allele frequency) (**Table 2**, **Figure 3**).

Thus, out of the twelve cfDNA samples, a total of four contained cf-mtDNA variants that were also present in the tumour tissue evaluated, as detected by either sequencing or dPCR. To put these results into perspective, we evaluated the levels of other blood-based cancer biomarkers in these samples. Tumour-specific mutations in nuclear-encoded driver genes were detected in the cfDNA (cf-nDNA) of three sera and three plasma samples (**Supplementary Table 5**), mostly at much higher allele frequencies than the detected tumour-specific cf-mtDNA variants (**Table 3**, **Supplementary Figures 2 and 3**). Also, the level of cancer-antigen was increased in the blood of P1, P2, P3 and P5 (≥ 30 kU/L) (**Table 3**, **Figure 3**, **Supplementary Figure 2**) and circulating tumour cells were detected in the blood of P6 and P8 (**Table 3**, **Supplementary Figure 3**).

Discussion

In this work, we show that there is extensive heterogeneity in mtDNA variants between tumour and tumour-adjacent normal tissue, and that tumour-specific cf-mtDNA variants are hardly detectable in the circulation of cancer patients.

To detect mtDNA variants, we used a SMRT sequencing approach to evaluate the whole mitochondrial genome [34] and included dPCR as an orthogonal method to evaluate a subset of the detected variants. The limit of detection for variants by our sequencing approach is mainly dependent on the sequencing depth at a position, with variants called based on at least five highly accurate single-molecule reads containing the variant at that position. We have recently shown that the SMRT sequencing approach is able to trace known low-frequent variants ≥ 0.1% allele frequency (sensitivity), and to reliably detect unknown variants ≥ 1.0% allele frequency (specificity) [34]. Therefore, we start with calling variants with at least 1.0% allele frequency in the evaluated samples (detection of unknown variants), and subsequently evaluate the presence of those called variants in the complete dataset (tracing of known variants). For some variants, we obtained a sequencing depth that allowed for tracing down to 0.04% allele frequency (**Supplementary Table 3**). Note that false positives due to random PCR errors are unlikely to be detected: we used minimally 100,000 input molecules (**Supplementary Table 1**) corresponding to a 0.001% allele frequency of random PCR errors (1 / 100,000), a high fidelity polymerase (error rate of ~10-7), and the number of PCR cycles was limited. To simultaneously evaluate the specificity of the method as well as the interference of nuclear insertions of mitochondrial origin (NUMTs), independent re-sequencing of four tissue specimens was performed after exonuclease-treatment of the DNA to specifically degrade

linear DNA and thus increase the circular mtDNA fraction. The latter is important since NUMTs can interfere with accurate variant detection due to their sequence similarity to mtDNA, and thus complicate investigation of mitochondrial heteroplasmy. Good concordance was observed: independent re-sequencing revealed that of the eighteen variants sixteen variants were confirmed. Two variants could not be confirmed: variant 6255G>A in P1 was detected first at 0.8% allele frequency but was below the limit of detection (< 0.9% allele frequency) in the re-sequenced specimen, and variant 9058A>G in P2 was not detected at first (limit of detection < 0.2% allele frequency) but was detected at 0.2% allele frequency in the re-sequenced specimen. These variants were called because they were present ≥1.0% allele frequency in the cfDNA samples of the corresponding patients. Both variants are not at putative known NUMT positions as evaluated by nucleotide BLAST. From this we concluded that the limiting factor in tracing variants by our approach is the sequencing depth, which influences the limit of detection.

The extensive heterogeneity we observe between tumour and tumour-adjacent normal tissue is in line with the observation by others that heteroplasmy patterns can differ in an allele-specific manner between tissues within an individual [3-6]. This is also evident from the high fraction of cf-mtDNA variants we detect that are of unknown tissue origin. Since the number of cell generations of epithelial tumour cells greatly exceeds that of non-tumour epithelial cells, it is likely that the intra-individual genetic drift observed between tissues within an individual likely also applies to tumour cells and their founder cells. Similarly, our observation that allele frequencies of tumour-specific variants are much higher than those observed in normal tissue, corresponds to the hypothesis that more cell generations equals more opportunity for either loss or expansion of a heteroplasmic mtDNA variant [35]. In line with this, only a low fraction of variants phase together in our work (19 of the 65 variant combinations, 30%), indicative that extensive heterogeneity is present among heteroplasmic variants within an individual. Remarkably, P2 shows an exceptional high number of heteroplasmic variants detected in only the normal specimen (n = 11), and the tumour of this patient also contained the highest number of heteroplasmic variants (n =5). Note that the primary tumour specimens of P6 and P8 did not contain tumour cells on the morphological level as evaluated by HE-stained slides (**Figure 1**, **Supplementary Table 1**), but on the molecular level the primary tumour specimen of P6 did contain tumour cells as evaluated by the presence of KRAS mutated nDNA (**Supplementary Table 5**). Uncertainty in estimating tumour cell percentage in HE-slides have been pointed out in literature [36], and fresh-frozen tissue sections are of lower morphological quality compared to formalin-fixed paraffin-embedded tissue sections, hampering tumour cell percentage estimation. Notable, whereas in P6 also the normal colon and normal liver appeared non-tumorous by morphological evaluation, mutated KRAS was present at the molecular level (Supplementary Table 5), indicating tumour cells are present in these

specimens as well. Thus, some of the variants defined as ubiquitous might actually be tumour-specific in P6 – especially variant 2305T>C. Additionally, it is interesting to see in this case that even variants at high allele frequency can be present on different mtDNA molecules (1924T>C and 2305T>C in P6, **Supplementary Table 4**).

With regard to the number of tumour-specific mtDNA variants per tumour, we observe a higher number compared to other studies. Specifically, other studies using large sample sizes showed that 75% of the breast cancer patients and 60% of the colon cancer patients harbour at least one mtDNA variant in their primary tumour based on massive parallel sequencing studies [7-9], whereas we observe in 100% (8/8) of the patients at least one tumour-specific mtDNA variant. Those studies applied a threshold on variant allele frequency between ≥ 3% and ≥ 15% allele frequency. We used an initial threshold at ≥ 1.0% and no threshold on allele frequency when the variant was called within another sample of the same patient. Based on our previous work, it is unlikely that this is due to false positive calls, since those appeared ≤ 1.0% variant allele frequency [34]. The higher number may thus be due to the higher sensitivity of our SMRT sequencing approach, or statistical co-incidence due to the relatively small series we analysed. Noteworthy is that whereas other studies have reported on (near-)homoplasmic tumour-specific mtDNA variants [7-9], we only observed heteroplasmic tumour-specific mtDNA variants < 50% allele frequency. This is likely due to the non-tumour cells present in our specimens (i.e. infiltrating immune cells or tissue of origin): it would not be possible to reach 100% allele frequency for a tumour-specific mtDNA variant since non-tumour mtDNA will be present as well.

Importantly, despite the presence of tumour-specific mtDNA variants for each tumour tissue analysed, we were unable to detect the majority of these variants in blood-circulating cfDNA (83% not detected). The few cf-mtDNA variants we detected were present at extremely low heteroplasmy levels (P1 and P7) or questionable in their true tumour-specific nature given the low heteroplasmy levels in the tumour tissue (P2 and P8). Specifically, in P2 variant 9058A>G was detected at 1.1% allele frequency in the cfDNA whereas it is present at 0.2% allele frequency in one of the two replicates of the primary tumour. In P2, the primary tumour also contained tumour-specific variants between 2% and 10% allele frequency (2724G>C, 10657T>C, 11040T>C and 16141A>G, **Figure** 1) but those were not detected as cfDNA in the circulation (**Figure 1**) (**Supplementary Table 3**). This would mean that – since the primary tumour was not *in situ* when blood was drawn – the metastases in this patient originated from a clone that contained only 9058A>G but not the other variants present in the primary tumour. Also, in P8 variant 16183A>C present at 0.7% allele frequency in the liver metastasis was detected at 3.8% allele frequency in the cfDNA. In this patient, the liver metastasis was still *in situ* when blood was drawn, but the tumour-specific variants present at 2% and 14% allele

frequency (8102G>A and 9196G>A, **Figure 1**) are not detected as cfDNA (**Figure 1**) (**Supplementary Table 3**). When comparing the detection of tumour-specific cf-mtDNA with other blood-based markers – including CTCs, cancer-antigens and mutations in cf-nDNA – the latter outperform cf-mtDNA since these could be detected in nearly all blood samples (**Table 3**).

Our hypothesis was that mutated cf-mtDNA would be easily detectable in the circulation due to high stability (if circular) and high copy number per cell, and thus would require less sensitive methods to detect them. However, the fact that nuclear-encoded mutations are detected at much higher allele frequencies in the cfDNA could indicate that either more mtDNA from non-tumour cells is released in the circulation, or that the cf-mtDNA has a higher turnover than its nuclear counterpart rendering it undetectable by the techniques we applied. Because the mtDNA content is variable per cell, it could be that tissues (non-tumour) with higher mtDNA content than the tumour are shedding DNA into the circulatory system. This could in part be the tumour-adjacent normal tissue: three-quarter of the breast tumours harbour a reduction in tumour mtDNA content compared to adjacent normal mammary tissue [2, 37-42], whereas approximately half of the colorectal tumours have a reduction in tumour mtDNA content when compared to adjacent normal colon or rectum tissue [2, 43-47]. Another likely source of non-tumour cf-mtDNA in blood is thrombocytes, which do not contain a nucleus but do contain mitochondria. Before being frozen, our plasma samples were obtained via a centrifugation force that should be enough to remove 90% of the thrombocytes [48]. After thawing, an additional centrifugation step was applied to obtain thrombocyte-poor plasma, but it could be that part of the thrombocytes had been damaged by freezing-thawing and thus released their mtDNA already into the plasma. Especially when comparing the mtDNA content of serum and plasma (order of hundred versus order of thousands, **Supplementary Table 1**), it seems evident that the capture of thrombocytes in the fibrin clot of serum results in less release of their mtDNA in the blood-derivate. Another probability is that cf-mtDNA is so severely fragmented that it is not detectable by our applied methods, which require DNA of at least 108 base pairs in length for dPCR and of at least 1,700 base pairs in length for SMRT sequencing (see amplicon sizes in **Supplementary Table 2**). A study on the physical characteristics of plasma-derived cf-mtDNA indicates that the majority is associated with particles between 5 μm and 0.45 μm in size, since filtering the plasma reduces the amount of cf-mtDNA whereas cf-nDNA is retained (note that the diameter and length of mitochondria range from resp. 0.5 to 1 μm and 5 to 10 μm) [14]. Treatment of serum-derived cfDNA with an exonuclease that digests linear DNA but leaves circular DNA intact, resulted in undetectable levels of nDNA, whereas mtDNA concentrations reduced 5-10x (unpublished data), indicative that at least a fraction of the cf-mtDNA is in its circular form within serum. Other studies, based on whole genome

sequencing of plasma-derived cfDNA without prior fragmentation of the DNA (thus, only short fragments smaller than ~ 600 base pairs are efficiently sequenced), indicate that part of cf-mtDNA is severely fragmented [15-18]. In our own hands, such a whole genome sequencing approach applied to plasma-derived cfDNA of cancer patients resulted in median 0.85x coverage (range 0.35-1.73x) of the mitochondrial contig, whereas the nuclear DNA was covered by ~1x (unpublished data). Taking into account the mtDNA content of cells (in the order of thousands in our plasma samples, **Supplementary Table 2**), this indicates that the largest proportion of cf-mtDNA is not sequenced by such a whole genome sequencing approach. This is in line with observations in hepatocellular cancer patients [16] and lung transplant recipients [17], where the fractional concentration of sequenced plasma cf-mtDNA was lower than expected. In the study on lung transplant recipients [17] and in another study on sepsis patients [18], the use of sequencing protocols that also include smaller DNA fragments (40 - 100 base pairs) increased the fraction of cf-mtDNA reads by 8- to 15-fold, but still cannot fully explain the low abundance of mtDNA in those experiments. Thus, it seems that a large proportion of the mtDNA is not sequenced in these studies, likely due to the fact that intact mtDNA is not efficiently sequenced during such approaches. It must be kept in mind that those results apply to plasma and not necessarily to serum. Another possibility why cf-mtDNA was rarely detectable might not be related to physical characteristics, but due to genetic drift: the tumour-specific variants present in the primary tumour are not present in the (micro-) metastases anymore. Especially when the elapsed time is large (e.g. > 40 weeks in P2, > 240 weeks P3 and > 400 weeks P5) it might be possible that the genetic make-up of the mtDNA in the tumours has changed, similar to the heterogeneity we observed between the tumour and tumour-adjacent normal tissue. Given the heterogeneity in mtDNA variants between tissues within an individual [3-6], it is not possible to evaluate all cf-mtDNA variants present in the circulation of a patient as tumour-specific ones, as illustrated by the number of cf-mtDNA variants from tumour-adjacent and of unknown origin in our study. Noteworthy, it could be that the total number of heteroplasmic cf-mtDNA variants present in the circulation is increased for (advanced) cancer patients, but this potentially is also the case for a patient affected with other morbidities (e.g. liver cirrhosis resulting in liver-specific cf-mtDNA variants, or colitis resulting in colon-specific cf-mtDNA variants).

Our results demonstrate that extensive mtDNA heterogeneity is evident within an individual. We conclude that there is limited value in tracing tumour-specific cf-mtDNA variants as a blood-circulating biomarker with the current methods available.

Materials and methods

Patient selection and sampling

We used material from our bio-bank at the department of Medical Oncology of the Erasmus MC Cancer Institute, Rotterdam, the Netherlands. Patient selection was based on availability of a frozen blood-derivate (plasma or serum) to obtain cfDNA and fresh frozen resection material of tumour tissue (primary or metastasis). For all except one case, fresh frozen material of normal tissue originating from the same resection material was available. Blood sampling was done in either serum separation tubes according to routine procedures in our hospital, or in EDTA tubes followed by cell separation within 24 hours after blood draw (10 minutes at 800 x g). Obtained serum or plasma samples had been stored at -80°C until use. After thawing, plasma samples underwent additional sedimentation at 1020 x g for 10 minutes at 4°C, of which the supernatant was used. Use of the patient material was approved by the medical ethics committee of the Erasmus MC (MEC 02.953 and MEC 06.089) and conducted in accordance to the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands.

DNA extraction

For fresh frozen tissue specimens, a DNA extraction method that enriches for mtDNA was performed as described before [34]. Briefly, 20 cryosections of 30 µm (average input of 30 mg tissue, range of 16 - 59 mg) per specimen were lysed to solubilize cellular membrane and release all cellular compartments (10 minutes, 1 mL of 0.5x TBE containing 0.5% (v/v) Triton X-100). Cell nuclei were removed (10 minutes 1020 x g) and DNA was extracted from the remaining supernatant using the QIAamp Circulating Nucleic Acid Kit (*Qiagen*) according to the suppliers' protocol. For four specimens, an additional sample was obtained by an independent DNA extraction (as described above) with subsequent enzymatic degradation for linear DNA as described before [34]. Briefly, those DNA extracts (max. 250 ng) were incubated with ATP-dependent exonuclease PlasmidSafe (*Epicentre*) (40 Units, 3 hours at 37°C), heat-inactivated (30 minutes 70°C) and purified (ethanol precipitation, 70% ethanol). For some frozen tissue specimens, DNA extracts were already available in our bio-bank, and had been obtained by either the PureLink Genomic DNA kit (*Invitrogen*) or the DNeasy Tissue Kit (*Qiagen*) as described by the supplier. For each tissue sample, 5 µm sections were obtained on microscopy slides and haematoxylin and eosin (HE)-stained to estimate the percentage of tumour cells within the sections used for DNA extraction. For the blood-derivates, after thawing at 4°C, DNA was extracted using the QIAamp Circulating Nucleic Acid Kit (*Qiagen*) according to the suppliers' protocol. Serum input ranged from 450 to 500 µL and plasma supernatant input was 1000 µL. Specifications for each sample are provided in **Supplementary Table 1**.

DNA quantification and mtDNA purity assessment

All DNA extracts were quantified using the Qubit dsDNA HS assay kit (*Life Technologies*) according to the suppliers' protocol. Purity of mtDNA was measured in duplicate runs of a multiplex qPCR assay targeting a nuclear and a mitochondrial encoded gene, to calculate the ratio of mtDNA molecules opposed to nDNA molecules by the relative quantitation method (2^ΔCq) as described before [49].

SMRT sequencing

SMRT sequencing was performed as described before [34]. Briefly, amplicons covering the complete mtDNA were generated by singleplex (tissue DNA) or multiplex (cfDNA) PCR reaction with initial denaturation for 3 minutes at 98°C, 15 or 18 cycles of a three-step PCR with 10 seconds denaturation (98°C), 30 seconds annealing (67°C) and 90 seconds extension (72°C), followed by a final extension (72°C) for 5 minutes. DNA input was set to contain at least 100,000 but maximally 50,000,000 copies mtDNA/reaction, based on the mtDNA content. Each 50 µL reaction contained DNA and 1 unit of Hot-Start Q5 High Fidelity DNA polymerase (*NEB*) in 1x Q5 reaction buffer, 200 µM dNTPs and 0.5 µM of 5'-M13 tailed forward and reverse primer (**Supplementary Table 2**). Specificity of the generated products was confirmed using microchip electrophoresis (DNA-12000 reagent kit, *Shimadzu*). Amplicons were equimolar pooled per sample and purified using AMPure PB paramagnetic beads (*Pacific Biosciences*) with a 0.6 beads:sample ratio according to the SMRTbell Template Prep Kit protocol and eluted in 10 mM Tris-HCl pH 8.5. The 5'-M13 universal sequence tail of the primers allowed barcoding of each sample by performing 5 amplification cycles of the three-step PCR as described above but with an annealing temperature of 58°C. Specificity of the generated products was confirmed using microchip electrophoresis (BioAnalyzer High Sensitivity DNA kit, *Agilent*). A final mix of barcoded fragments was obtained by pooling of multiple samples and subsequently purified using AMPure PB paramagnetic beads with a 0.6 beads:sample ratio. Concentration of the final mix was determined using the Qubit dsDNA HS assay kit, and SMRTbell library was generated according to the Amplicon Template Preparation and Sequencing guide (*Pacific Biosciences*). Sequencing was performed on Pacific Biosciences RSII with P6-C4 sequencing chemistry and 360 minutes movie-time. For the tissue samples, a total of 15 SMRT cells were used to reach a read depth estimated at 600x per sample, and for the cfDNA samples a total of 28 SMRT cells were used to reach a read depth estimated at 3000x per sample. Specifications for each sample are provided in **Supplementary Table 1**.

Digital PCR

Digital PCR (dPCR) was performed on the Quantstudio 3D digital PCR system (*Thermo Fisher*) according to the supplier's protocol. Detection of KRAS p.G12D, KRAS p.G12V, TP53 p.R248Q, TP53 p.R273H and PIK3CA p.H1047R was done by validated Taqman SNP genotyping assays (*Thermo Fisher*). Detection of mtDNA variants 664G>A, 6255G>A, 1924T>C and 2305T>C was done by custom assays (**Supplementary Table 2**) and carried out with an adaption to the DNA input due to high mtDNA copy number (set to contain at least 1E+3 but maximally 2E+4 copies mtDNA/reaction based on mtDNA content). Reactions contained DNA in 1x dPCR mastermix v2, 0.9 μM of each primer and 0.2 μM of each probe. After initial denaturation for 10 minutes at 96°C, the 40-cycle two-step PCR was performed at 30 seconds denaturation (98°C) and 120 seconds annealing/extension (56°C), and followed by a final 2 minute extension (56°C). To calculate allele frequency of the alternative variant, the threshold for signal dots was set to at least two dots per dye. For samples where the variant was not detected, the limit of detection was calculated based on the total number of positive wildtype dots and two mutant signal dots (e.g. 5000 wildtype dots would correspond to a detection limit of 2 / 5000 = 0.04%).

Bioinformatics

RS bax.h5 files were converted to Sequel BAM files, of which circular consensus reads (CCS) were generated using the CCS2 algorithm, and attributed to each sample using the sample-specific barcode [50]. Next, a minimum quality threshold of 99% and at least five passes of the SMRTbell were applied to select for highly accurate single-molecule reads. Selected CCS reads were trimmed (Cutadapt [51] for primers-tails) and subsequently aligned against a reference sequence (BWA-MEM parameters -k17 -W40 -r10 -A1 -B1 -O1 -E1 -L0 [52]). As reference sequence, we used an extended version of rCRS to compensate for mapping bias due to circularity of the mitochondrial genome. Positions alternative to the reference sequence in pileup files (Bioconductor Rsamtools 1.26.2 pileup function with pileupParam min_base_quality=30, min_mapq=0, min_nucleotide_depth=5, min_minor_allele_depth=0, distinguish_strands=TRUE, distinguish_nucleotides=TRUE, ignore_query_Ns=TRUE, include_deletions=FALSE, include_insertions=FALSE) were converted back to rCRS positions and used for analyses. Allele frequency was calculated based on the alternative variant (alternative reads / total reads). First, all homoplasmic and high heteroplasmic alternative variants (> 50% allele frequency) were used for haplotyping (HaploGrep2 v2.1.0) to determine patient-specificity for each sample (**Supplementary Figure 1**). Then, initial variant selection was performed on the pileup files using a threshold of 1-99% alternative allele frequency. Those variants were manually inspected in Integrative

Genomics Viewer (IGV, *Broad Institute*) [53] to exclude mapping artefacts (as elaborated on in [34]). Of the remaining variants, their presence within the initial pileup file was determined in all examined samples of that patient to generate a final list of detected variants (**Supplementary Table 3**). Also, the detected heteroplasmic variants were used in a nucleotide BLAST against the human reference sequence (NCBI's nucleotide web blast, https://blast.ncbi.nlm.nih.gov) with the surrounding reference sequence (30 bases 5' and 30 bases 3') to uncover potential NUMT events, but none were recovered. For the samples in which variants were not called within the final list, limit of detection at that position was calculated based on the read depth at that position and an alternative variant read depth of 5 (e.g. a position with 5000x read depth would correspond to a detection limit of 5 / 5000 = 0.1%).

Supplementary data

Supplementary data for this article are available online at Neoplasia (https://www.journals.elsevier.com/neoplasia/).

References

1. Wachsmuth, M., et al., *Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number.* PLoS Genet, 2016. **12**(3): e1005939.
2. Reznik, E., et al., *Mitochondrial DNA copy number variation across human cancers.* Elife, 2016. **5**: e10769.
3. Samuels, D.C., et al., *Recurrent tissue-specific mtDNA mutations are common in humans.* PLoS Genet, 2013. **9**(11): e1003929.
4. He, Y., et al., *Heteroplasmic mitochondrial DNA mutations in normal and tumour cells.* Nature, 2010. **464**(7288): p. 610-614.
5. Li, M., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations.* Proc Natl Acad Sci U S A, 2015. **112**(8): p. 2491-2496.
6. Calloway, C.D., et al., *The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age.* Am J Hum Genet, 2000. **66**(4): p. 1384-1397.
7. Ju, Y.S., et al., *Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer.* Elife, 2014. **3**: e02935.
8. Larman, T.C., et al., *Spectrum of somatic mitochondrial mutations in five cancers.* Proc Natl Acad Sci U S A, 2012. **109**(35): p. 14087-14091.
9. Stewart, J.B., et al., *Simultaneous DNA and RNA mapping of somatic mitochondrial mutations across diverse human cancers.* PLoS Genet, 2015. **11**(6): e1005333.
10. Brown, W.M., M. George, Jr., and A.C. Wilson, *Rapid evolution of animal mitochondrial DNA.* Proc Natl Acad Sci U S A, 1979. **76**(4): p. 1967-1971.
11. Diaz, L.A., Jr. and A. Bardelli, *Liquid biopsies: genotyping circulating tumor DNA.* J Clin Oncol, 2014. **32**(6): p. 579-586.
12. Gilkerson, R., et al., *The mitochondrial nucleoid: integrating mitochondrial DNA into cellular homeostasis.* Cold Spring Harb Perspect Biol, 2013. **5**(5): a011080.
13. Murgia, M., et al., *Mitochondrial DNA is not fragmented during apoptosis.* J Biol Chem, 1992. **267**(16): p. 10939-10941.
14. Chiu, R.W., et al., *Quantitative analysis of circulating mitochondrial DNA in plasma.* Clin Chem, 2003. **49**(5): p. 719-726.
15. Chandrananda, D., N.P. Thorne, and M. Bahlo, *High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA.* BMC Med Genomics, 2015. **8**: 29.
16. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.* Proc Natl Acad Sci U S A, 2015. **112**(11): E1317-25.
17. Burnham, P., et al., *Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma.* Sci Rep, 2016. **6**: 27859.
18. Zhang, R., et al., *Very short mitochondrial DNA fragments and heteroplasmy in human plasma.* Sci Rep, 2016. **6**: 36097.
19. Zachariah, R.R., et al., *Levels of circulating cell-free nuclear and mitochondrial DNA in benign and malignant ovarian tumors.* Obstet Gynecol, 2008. **112**(4): p. 843-850.
20. Cheng, C., et al., *Quantification of circulating cell-free DNA in the plasma of cancer patients during radiation therapy.* Cancer Sci, 2009. **100**(2): p. 303-309.
21. Mead, R., et al., *Circulating tumour markers can define patients with normal colons, benign polyps, and cancers.* Br J Cancer, 2011. **105**(2): p. 239-245.
22. Mehra, N., et al., *Circulating mitochondrial nucleic acids have prognostic value for survival in patients with advanced prostate cancer.* Clin Cancer Res, 2007. **13**(2): p. 421-426.
23. Mahmoud, E.H., et al., *Plasma circulating cell-free nuclear and mitochondrial DNA as potential biomarkers in the peripheral blood of breast cancer patients.* Asian Pac J Cancer Prev, 2015. **16**(18): p. 8299-8305.
24. Jeronimo, C., et al., *Mitochondrial mutations in early stage prostate cancer and bodily fluids.* Oncogene, 2001. **20**(37): p. 5195-5198.

25. Uzawa, K., et al., *Circulating tumor-derived mutant mitochondrial DNA: a predictive biomarker of clinical prognosis in human squamous cell carcinoma.* Oncotarget, 2012. **3**(7): p. 670-677.
26. Takeuchi, H., A. Fujimoto, and D.S. Hoon, *Detection of mitochondrial DNA alterations in plasma of malignant melanoma patients.* Ann N Y Acad Sci, 2004. **1022**: p. 50-54.
27. Hibi, K., et al., *Detection of mitochondrial DNA alterations in primary tumors and corresponding serum of colorectal cancer patients.* Int J Cancer, 2001. **94**(3): p. 429-431.
28. Losanoff, J.E., et al., *Can mitochondrial DNA mutations in circulating white blood cells and serum be used to detect breast cancer?* Breast, 2008. **17**(5): p. 540-542.
29. Okochi, O., et al., *Detection of mitochondrial DNA alterations in the serum of hepatocellular carcinoma patients.* Clin Cancer Res, 2002. **8**(9): p. 2875-2878.
30. Fliss, M.S., et al., *Facile detection of mitochondrial DNA mutations in tumors and bodily fluids.* Science, 2000. **287**(5460): p. 2017-2019.
31. Zhu, W., et al., *Mitochondrial DNA mutations in breast cancer tissue and in matched nipple aspirate fluid.* Carcinogenesis, 2005. **26**(1): p. 145-152.
32. Wong, L.J., et al., *Detection of mitochondrial DNA mutations in the tumor and cerebrospinal fluid of medulloblastoma patients.* Cancer Res, 2003. **63**(14): p. 3866-3871.
33. Duberow, D.P., et al., *High-performance detection of somatic D-loop mutation in urothelial cell carcinoma patients by polymorphism ratio sequencing.* J Mol Med (Berl), 2016. **94**(9): p. 1015-1024.
34. Weerts, M.J.A., et al., *Sensitive detection of mitochondrial DNA variants for analysis of mitochondrial DNA-enriched extracts from frozen tumor tissue.* Scientific Reports, 2018. **8**(1): 2261.
35. Coller, H.A., et al., *High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection.* Nat Genet, 2001. **28**(2): p. 147-150.
36. Smits, A.J., et al., *The estimation of tumor cell percentage for molecular testing by pathologists is not accurate.* Mod Pathol, 2014. **27**(2): p. 168-174.
37. Mambo, E., et al., *Tumor-specific changes in mtDNA content in human cancer.* Int J Cancer, 2005. **116**(6): p. 920-924.
38. Yu, M., et al., *Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients.* IUBMB Life, 2007. **59**(7): p. 450-457.
39. Tseng, L.M., et al., *Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer.* Genes Chromosomes Cancer, 2006. **45**(7): p. 629-638.
40. Fan, A.X., et al., *Mitochondrial DNA content in paired normal and cancerous breast tissue samples from patients with breast cancer.* J Cancer Res Clin Oncol, 2009. **135**(8): p. 983-989.
41. Barekati, Z., et al., *Methylation profile of TP53 regulatory pathway and mtDNA alterations in breast cancer patients lacking TP53 mutations.* Hum Mol Genet, 2010. **19**(15): p. 2936-2946.
42. McMahon, S. and T. LaFramboise, *Mutational patterns in the breast cancer mitochondrial genome, with clinical correlates.* Carcinogenesis, 2014. **35**(5): p. 1046-1054.
43. Mohideen, A.M., et al., *Mitochondrial DNA polymorphisms, its copy number change and outcome in colorectal cancer.* BMC Res Notes, 2015. **8**: 272.
44. Feng, S., et al., *Correlation between increased copy number of mitochondrial DNA and clinicopathological stage in colorectal cancer.* Oncol Lett, 2011. **2**(5): p. 899-903.
45. Chen, T., et al., *The mitochondrial DNA 4,977-bp deletion and its implication in copy number alteration in colorectal cancer.* BMC Med Genet, 2011. **12**: 8.
46. Gao, J., et al., *De-methylation of displacement loop of mitochondrial DNA is associated with increased mitochondrial copy number and nicotinamide adenine dinucleotide subunit 2 expression in colorectal cancer.* Mol Med Rep, 2015. **12**(5): p. 7033-7038.
47. Cui, H., et al., *Association of decreased mitochondrial DNA content with the progression of colorectal cancer.* BMC Cancer, 2013. **13**: 110.
48. Perez, A.G., et al., *Relevant aspects of centrifugation step in the preparation of platelet-rich plasma.* ISRN Hematol, 2014. **2014**: 176060.

49. Weerts, M.J.A., et al., *Mitochondrial DNA content in breast cancer: Impact on in vitro and in vivo phenotype and patient prognosis.* Oncotarget, 2016. **7**(20): p. 29166-29176.
50. Anvar, S.Y., et al., *TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes.* Bioinformatics, 2014. **30**(12): p. 1651-1659.
51. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): 10-12.
52. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-595.
53. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-26.

CHAPTER 9


ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
Thr
Pro
CYB
Glu
Asp
Ser
CO2
Lys
ATP8
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Leu

Summary / Samenvatting

The past decades, somatic alterations have been characterized in the tumour's DNA, increasing our knowledge of human cancer and paving the way toward new diagnostics and treatments. More recently, blood-circulating cell-free DNA (cfDNA) as cancer biomarker is extensively studied to uncover tumour-specific alterations in a minimal invasive manner, and might proof fruitful in early detection, providing prognostic or predictive information guiding therapy-decision making, or to monitor treatment response or the burden of residual disease. Nevertheless, almost oblivion in these endeavours is the human mitochondrial DNA (mtDNA), a gene-dense circular entity present as numerous copies within our cells. As described in **Chapter 1**, in this thesis we extended the exploration on the value of somatic mtDNA alterations, and cfDNA, as a cancer biomarker.

In **Chapter** 2, we aimed to clarify the link between mtDNA content and a mesenchymal phenotype and its relation to prognosis of breast cancer patients. The number of mtDNA molecules per cell was quantified in 42 breast cancer cell lines and 207 primary breast tumour specimens. We did not find evidence between mtDNA content and the mesenchymal phenotype, but low mtDNA was associated with a poorer distant metastasis free survival. We concluded that mtDNA content might provide meaningful prognostic value for distant metastasis in breast cancer.

In **Chapter 3**, we aimed to explore whether low levels of mtDNA content in the primary tumour could predict better outcome for breast cancer patients receiving anthracycline-based therapies. The number of mtDNA molecules per cell was quantified in 295 primary breast tumour specimens. Low mtDNA was associated with better distant metastasis-free survival and progression-free survival when patients were treated with anthracyline-based chemotherapy, but not when treated with methotrexate-based chemotherapy. We concluded that mtDNA content in primary breast tumours may be exploited by guiding chemotherapeutic regimen decision-making.

In **Chapter 4**, we aimed to further understand the genomic changes and expression of the mtDNA within breast cancer. RNA sequencing data of 344 primary breast tumour specimens was analysed. The number of somatic variants within mtRNA was not associated with the mutational processes impacting the nuclear genome, but positively correlated with age at diagnosis. Also, mitochondrial expression was related to ER status or the primary tumour. We concluded that there is a large heterogeneity in somatic mutations of the mtDNA within primary breast tumours, and differences in mitochondrial expression among breast cancer subtypes.

In **Chapter 5**, we aimed to detect tumour-specific mutations in only minute amounts of serum-derived cfDNA by using a targeted next generation sequencing (NGS) approach. The nuclear DNA of primary tumour tissue (fresh frozen), tumour-adjacent normal tissue (formalin-fixed paraffin embedded), and three consecutive serum samples (frozen) from 10 cancer patients was sequenced by a targeted-sequencing approach and a workflow

developed. We concluded that, our workflow was able to detect variants traceable as circulating tumour DNA (ctDNA) in minimal amounts of sera of cancer patients.

In **Chapter 6**, we aimed to develop a procedure to detect low-frequent single-nucleotide mtDNA-specific variants. Four methods to extract mtDNA as pure as possible from frozen tumour tissue, and three methods for low-frequent variant detection were evaluated. We concluded that our sensitive procedure to detect low-frequent single-nucleotide mtDNA variants from frozen tumour tissue is based on extraction of DNA from cytosol fractions followed by exonuclease treatment to obtain high mtDNA purity, and subsequent SMRT sequencing for (*de novo*) detection and allelic phasing of variants.

In **Chapter 7**, we aimed to explore the potential of tumour-specific cf-mtDNA variants as cancer marker in the blood of cancer patients. The entire mtDNA of primary tumour and/or metastatic sites, tumour-adjacent normal tissue, and cfDNA originating from 8 cancer patients was sequenced by the in Chapter 6 developed procedure. Extensive heterogeneity was observed among the heteroplasmic mtDNA variants present within each individual, and the few tumour-specific mtDNA variants detected in cfDNA were present at much lower allele frequencies as nuclear-encoded somatic mutations. We concluded that there is limited value in tracing tumour-specific mtDNA variants in blood-circulating cfDNA with the current methods available.

From this thesis, as described in **Chapter 8**, it can be concluded that the mtDNA content in the primary breast tumour is not associated with any of the traditional clinicopathological markers, mtDNA expression differs among breast cancer subtypes, and patients diagnosed at a higher age harbour more somatic mtDNA variants in their primary tumour. Also, there is large heterogeneity in somatic mtDNA variants within primary breast tumours. The absence of mutational hotspots in mtDNA, the extensive heterogeneity in heteroplasmic mtDNA variants within an individual, and the low occurrence of tumour-specific mtDNA variants in cfDNA, makes tracing tumour-specific cf-mtDNA variants as tumour biomarker in breast cancer a too-tailored approach. Also, a low level of mtDNA in the primary tumour indicates a more aggressive cancer, but at the same time also more susceptibility for anthracycline-based regimen. Hypothetically, this also applies other chemotherapeutic regimen known to induce high oxidative stress in the mitochondria. The susceptibility of non-tumour cells to treatments that induce (severe) oxidative stress and their contribution to cancer-related fatigue deserves to be studied. In the future, a complete picture on the role of mitochondrial variation in tumours should be obtained by studying primary tumour and metastatic sites for the interactions between the nuclear genome and the mitochondrial genome, taking into account both germline and somatic variation, but also tissue distribution and tissue dependence on oxidative phosphorylation.

De afgelopen decennia is onze kennis van kanker bij de mens toegenomen, waarbij de meest voorkomende verkregen veranderingen in het DNA van tumoren in kaart zijn gebracht. Dit heeft deuren geopend voor nieuwe mogelijkheden binnen de diagnostiek en behandeling van kanker. Om deze tumor specifieke veranderingen op een minimaal invasieve manier te kunnen detecteren wordt op dit moment bloed-circulerend cel-vrij DNA (cfDNA) uitgebreid bestudeerd als bio-indicator. De verwachting is dat cfDNA waardevol kan zijn bij vroege detectie van kanker, het verschaffen van prognostische of voorspellende informatie voor therapiebeslissingen, het kunnen volgen van de effectiviteit van een antikankerbehandeling, en bij het bepalen of er nog resterende ziekte aanwezig is. In deze inspanningen wordt het mitochondriële DNA (mtDNA), een cirkelvormig DNA-molecuul aanwezig in talloze kopieën binnen onze cellen, vaak buiten beschouwing gelaten. Zoals beschreven in **Hoofdstuk 1**, hebben we in dit proefschrift getracht om de waarde van het mtDNA en cfDNA als een bio-indicator voor kanker uit te breiden.

In **Hoofdstuk 2** hebben we het verband tussen de hoeveelheid mtDNA in tumorcellen en het mesenchymale fenotype, en de relatie tot de prognose van borstkankerpatiënten onderzocht. Het aantal mtDNA-moleculen werd gekwantificeerd in 42 borstkankercellijnen en 207 primaire-borsttumormonsters. We vonden geen aanwijzingen voor een relatie tussen de hoeveelheid mtDNA en het mesenchymale fenotype, maar een lage hoeveelheid mtDNA was geassocieerd met een slechtere metastase-vrije overleving. We concludeerden dat de hoeveelheid mtDNA in de primaire tumor van prognostische waarde zou kunnen zijn voor het voorspellen van uitzaaiingen op afstand bij borstkanker.

In **Hoofdstuk 3** wilden we onderzoeken of een lage hoeveelheid mtDNA in de primaire tumor een betere uitkomst zou kunnen voorspellen voor borstkankerpatiënten die behandeld worden met anthracycline-gebaseerde therapieën. Het aantal mtDNA-moleculen werd gekwantificeerd in 295 primaire-borsttumormonsters. Een lage hoeveelheid mtDNA werd geassocieerd met een betere metastasevrije overleving en progressievrije overleving wanneer patiënten werden behandeld met anthacylinegebaseerde chemotherapie, maar dit was niet het geval wanneer ze werden behandeld met methotrexaatgebaseerde chemotherapie. We concludeerden dat de hoeveelheid mtDNA in primaire borsttumoren zou kunnen worden benut om de besluitvorming rond de chemotherapiekeuze te begeleiden.

In **Hoofdstuk 4** wilden we veranderingen in en expressie van het mtDNA bij borstkanker verder uitdiepen. RNA-sequentiegegevens van 344 primaire-borsttumormonsters werden geanalyseerd. Het aantal verkregen varianten in het mtRNA was niet geassocieerd met de mutatieprocessen die invloed hadden op verkregen veranderingen in DNA uit de celkern, maar was wel positief gecorreleerd met de leeftijd waarop de kanker gediagnostiseerd werd. De mitochondriale expressie was gerelateerd aan de oestrogeenreceptorstatus van de primaire tumor. We concludeerden dat er een grote heterogeniteit bestaat in ver-

kregen veranderingen in het mtDNA van primaire borsttumoren, en dat er verschillen zijn in mitochondriële expressie tussen subtypen van borstkanker.

In **Hoofdstuk 5** hebben we ernaar gestreefd om tumorspecifieke veranderingen te detecteren in cfDNA uit slechts zeer kleine hoeveelheden bloedserum, door middel van een gerichte next generation sequencing-benadering (NGS). Het kern-DNA van primair-tumorweefsel, tumor-aangrenzend normaal weefsel, en drie opeenvolgende bloedserum-monsters van tien kankerpatiënten werden gericht gesequencet. We concludeerden dat onze werkwijze in staat was om tumorspecifieke veranderingen te traceren als bio-indicator in het cfDNA (verkregen uit minimale hoeveelheden bloedsera) van kankerpatiënten.

In **Hoofdstuk 6** wilden we een procedure ontwikkelen om laagfrequente veranderingen in mtDNA te detecteren. Vier methoden om mtDNA zo zuiver mogelijk uit tumorweefsel te extraheren en drie methoden voor laagfrequente-variantdetectie werden geëvalueerd. We kwamen tot een werkwijze om laagfrequente veranderingen in mtDNA detecteren door de extractie van DNA uit cytosolfracties van tumorweefsel gevolgd door enzymatische exonucleasebehandeling om een hoge mtDNA-zuiverheid te verkrijgen, en daaropvolgend SMRT-sequensen voor detectie en allelfasering van mtDNA-varianten.

In **Hoofdstuk 7** wilden we het potentieel van tumorspecifieke veranderingen in cf-mtDNA als kanker bio-indicator in het bloed van patiënten onderzoeken. Het volledige mtDNA van weefsel van de primaire tumor en/of uitzaaiingen, tumor-aangrenzend normaal weefsel en cfDNA afkomstig van acht kankerpatiënten werd gesequencet volgens de werkwijze ontwikkeld in Hoofdstuk 6. Grote heterogeniteit werd waargenomen tussen de mtDNA-varianten die aanwezig waren in elk individu. De weinige tumorspecifieke veranderingen in mtDNA die werden gedetecteerd in cfDNA waren aanwezig met veel lagere allelfrequenties dan de gedetecteerde verkregen veranderingen in het kern-DNA. We concludeerden dat er met de huidige beschikbare methoden beperkte waarde is in het traceren van tumorspecifieke veranderingen in mtDNA als bloedcirculerend cfDNA.

Zoals beschreven in **Hoofdstuk 8**, kan uit dit proefschrift worden geconcludeerd dat de hoeveelheid mtDNA in de primaire borsttumor niet is geassocieerd met een van de traditionele klinisch-pathologische bio-indicatoren, dat mtDNA-expressie verschilt tussen subtypen van borstkanker, en dat er een hoger aantal verkregen mtDNA-varianten aanwezig zijn in de borsttumoren van patiënten die gediagnosticeerd zijn op een hogere leeftijd. Er is een grote heterogeniteit in de verkregen mtDNA-varianten binnen primaire borsttumoren. De afwezigheid van mutatie-hotspots in mtDNA, de uitgebreide heterogeniteit in mtDNA-varianten binnen een individu, en het zelden voorkomen van tumorspecifieke mtDNA-varianten in cfDNA maken het opsporen van tumorspecifieke cf-mtDNA-varianten als bio-indicator bij borstkanker een niet-generaliseerbare benadering. Ook wijst een lage hoeveelheid mtDNA in de primaire borsttumor op een agressievere kanker, maar tegelijkertijd ook op meer gevoeligheid voor anthracycline-

gebaseerde chemotherapie. Hypothetisch zou dit ook van toepassing kunnen zijn op andere chemotherapeutische behandelingen waarvan bekend is dat ze hoge oxidatieve stress in de mitochondriën veroorzaken. De gevoeligheid van niet-tumorcellen voor behandelingen die (ernstige) oxidatieve stress veroorzaken en de potentiele bijdrage hiervan aan kankergerelateerde vermoeidheid verdienen het om te worden onderzocht. In de toekomst zou een vollediger beeld van de rol van veranderingen in het mtDNA in kanker moeten worden verkregen, door de primaire tumor én uitzaaiingen te bestuderen voor de interacties tussen het nucleaire genoom en het mitochondriële genoom. Hier moeten zowel kiemlijn- als verkregen variatie, maar ook de weefselverdeling en weefselafhankelijkheid van oxidatieve fosforylering in acht moet worden genomen.

ND1
Leu
Ile
Gln
Met
RNR2
RNR1
Val
Phe
ND2
Trp
Ala
Asn
Cys
Tyr
COI
CYB
Thr
Pro
Glu
Asp
Ser
CO2
ATP8
Lys
ATP6
ND6
CO3
ND4
ND5
Gly
ND3
Arg
ND4l
His
Leu
Arg

Appendices

List of publications

Marjolein JA Weerts, Marcel Smid, John A Foekens, Stefan Sleijfer, John WM Martens. Mitochondrial RNA Expression and Single Nucleotide Variants in Association with Clinical Parameters in Primary Breast Cancers. Cancers (10). 2018 Dec 9. doi: 10.3390/cancers10120500

Marjolein JA Weerts, Eveline C Timmermans, Anja van de Stolpe, Rolf HAM Vossen, Seyed Y Anvar, John A Foekens, Stefan Sleijfer, John WM Martens. Tumour-specific mitochondrial DNA variants are rarely detected in cell-free DNA. Neoplasia (20). 2018 May 26. doi: 10.1016/j.neo.2018.05.003

Marjolein JA Weerts, Eveline C Timmermans, Rolf HAM Vossen, Dianne van Strijp, Mirjam CGN Van den Hout–van Vroonhoven, Wilfred FJ van IJcken, Pieter-Jan van der Zaag, Seyed Y Anvar, Stefan Sleijfer, John WM Martens. Sensitive detection of mitochondrial DNA variants for analysis of mitochondrial DNA-enriched extracts from frozen tumour tissue. Scientific reports (8). 2018 Feb 02. doi: 10.1038/s41598-018-20623-7

Marjolein JA Weerts, Antoinette Hollestelle, Anieta M Sieuwerts, Marion E Meijer – van Gelder, John A Foekens, Stefan Sleijfer, John WM Martens. Low mitochondrial DNA content is associated with better outcome in breast cancer patients receiving anthracycline-based chemotherapy. Clinical Cancer Research (23). 2017 Aug. doi: 10.1158/1078-0432.CCR-17-0032

Marjolein JA Weerts*, Ronald van Marion*, Jean CA Helmijr, Corine M Beaufort, Niels MG Krol, Anita MAC Trapman-Jansen , Winand NM Dinjens, Stefan Sleijfer, Maurice PHM Jansen, John WM Martens. Somatic tumour mutations detected by targeted next generation sequencing in minute amounts of serum-derived cell-free DNA. Scientific Reports (7). 2017 May 18. doi: 10.1038/s41598-017-02388-7

Nick Beije, Jean C Helmijr, Marjolein JA Weerts, Corine M Beaufort, Matthew Wiggin, Andre Marziali, Cornelis Verhoef, Stefan Sleijfer, Maurice PHM Jansen, John WM Martens. Somatic mutation detection using various targeted detection assays in paired samples of circulating tumour DNA, primary tumour and metastases from patients undergoing resection of colorectal liver metastases. Molecular Oncology (10). 2016 Oct 10. doi: 10.1016/j.molonc.2016.10.001.

Maurice PHM Jansen, John WM Martens, Jean CA Helmijr, Corine M Beaufort, Ronald van Marion, Niels MG Krol, Kim Monkhorst, Anita MAC Trapman-Jansen, Marion E Meijer-van Gelder, Marjolein JA Weerts, Diana E Ramirez-Ardila, Hendrikus Jan Dubbink, John A Foekens, Stefan Sleijfer, Els MJJ Berns. Cell-free DNA mutations as biomarkers in breast cancer patients receiving tamoxifen. Oncotarget (7). 2016 May 30. doi: 10.18632/oncotarget.9727.

Marjolein JA Weerts, Anieta M Sieuwerts, Marcel Smid, Maxime P Look, John A Foekens, Stefan Sleijfer, John WM Martens. Mitochondrial DNA content in breast cancer: Impact on in vitro and in vivo phenotype and patient prognosis. Oncotarget (7). 2016 Apr 11. doi: 10.18632/oncotarget.8688.

Elisa Matas-Rico, Michiel van Veen, Kasia Kedzoria, Jan Koster, Daniela Leyton-Puig, Bas Molenaar, Marjolein JA Weerts, Ben NG Giepman, Anastassis Perrakis, Kees Jalink, Rogier Versteeg, Wouter H Moolenaar. GDE2 induces neuronal differentiation through glypican cleavage and is a marker of neuroblastoma outcome. Cancer Cell (30). 2016 Oct 10. doi: 10.1016/j.ccel.2016.08.016

Christopher J Daughney, Adrian Hetzer, Hannah T Heinrich, Peta-Gaye G Burnett, Marjolein Weerts, Hugh Morgan, Phil J Bremer, AJ McQuillan. Proton and cadmium adsorption by the archaeon Thermococcus zilligii: Generalising the contrast between thermophiles and mesophiles as sorbents. Chemical Geology. 2010 (273): 82-90. doi: 10.1016/j.chemgeo.2010.02.014

ND5
His
Leu
Ser
ND4
ND4L
Arg
ND3
Gly
CO3
ATP6
Lys
ATP8
Asp
Ser
CO2
Trp
Cys
Asn
Ala
Tyr
CO1
Gln
ND6
Pro
Thr
CYB
Phe
RNA1
Val
RNA2
ND2
Met
Ile
Leu
ND1