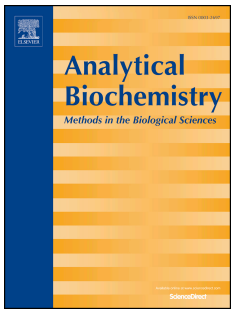
Array-based Dynamic Allele Specific Hybridization (Array-DASH): optimization-free microarray processing for multiple simultaneous genomic assays

DOI:
10.1016/j.ab.2021.114124

Document Version
Accepted author manuscript

Journal Pre-proof



Array-based Dynamic Allele Specific Hybridization (Array-DASH): optimization-free microarray processing for multiple simultaneous genomic assays

Spencer J. Gibson, Nathalie Zahra, Peter.J. Freeman, Caroline Howard, Owen Lancaster, Colin Veal, Maria Casadellà Fontdevila, Roger Paredes, Marc Noguera-Julian, Adrian Slater, Anthony J. Brookes

PII: S0003-2697(21)00025-7

DOI: https://doi.org/10.1016/j.ab.2021.114124

Reference: YABIO 114124

To appear in: *Analytical Biochemistry*

Received Date: 28 September 2020

Revised Date: 25 January 2021

Accepted Date: 26 January 2021

Please cite this article as: S.J. Gibson, N. Zahra, P.J. Freeman, C. Howard, O. Lancaster, C. Veal, M.C. Fontdevila, R. Paredes, M. Noguera-Julian, A. Slater, A.J. Brookes, Array-based Dynamic Allele Specific Hybridization (Array-DASH): optimization-free microarray processing for multiple simultaneous genomic assays, *Analytical Biochemistry*, https://doi.org/10.1016/j.ab.2021.114124.

Authors contribution.

Spencer J. Gibson (SG)

Development of the core ArrayDASH process including adaptation, development and optimisation of the ArrayDASH process using synthetic oligonucleotides, buffer optimisation, protocol development and optimisation, optimisation of array synthesis requirements (i.e. array manufacturer, array and spacer type) and LSCAN array design. ArrayDASH data interpretation including initial manual interpretation to check and refine the basic principles of DASH for use with ArrayDASH for both re-sequencing and genotyping applications. Manual interpretation of ArrayDASH data (either initially or as a check of data obtained) and input into the interpretation flow chart in figure 6. Assistance in the preparation of HIV targets and technical advice on the development of the PlantID and HIV arrays and the product purification and amplification processes. Assisted with the obtaining of the iNet grant that funded the PlantID work.

Nathalie Zahra. (NZ)

Development of the target amplification and purification process for HIV and ITS targets. Design of the HIV and Plant ITS arrays. Optimisation, performance and initial manual data extraction and results interpretation (base and genotyping calls only) of the HIV and ITS arrays. Technical advice and assistance with the development of the core ArrayDASH protocol.

Peter J. Freeman (PF)

Technical advice on the development of the core array DASH experimental methodology, particularly around suggestions to improve sensitivity and fidelity of PCR amplifications.

Caroline Howard (CH)

Sourcing of *Hypericum* material, amplification of the ITS regions from Hypericum samples. Sequencing of Hypericum species ITS and RBCL mitochondrial genome sequences and their alignments used in the design of the PlantID arrays by NZ.

Owen Lancaster (OL)

Development of the in-house Perl programs used to produce both re-sequencing and genotyping probes from the chosen target regions. Also, the development of an initial re-sequencing program (not used in the manuscript) automating calls from ArrayDASH data where there were no mixed alleles (i.e. samples were homozygous).

Colin Veal (CV)

General technical advice on the development of the core ArrayDASH process and initial suggestions for automating the analysis of resulting ArrayDASH data. Statistical assistance in comparing the ArrayDASH data to the HIV re-sequencing data.

Maria Casadellà Fontdevila (MCF), Roger Paredes (RP) and

Sourcing, amplification and shipment of the control and patient HIV samples. N.B. samples were amplified at the Spanish facility and only PCR products from the target regions of the reverse transcriptase and protease genes were shipped (in accordance with shipping regulations) to the UK. Next Generation Sequencing of the HIV samples provided to UK team.

Marc Noguera-Julian (MNJ)

Oversight and leadership of the activities of MCF and RP

Adrian Slater (AS)

Oversight and leadership of the activity of CH including her Doctoral thesis which formed the basis of the PlantID work. Obtained the iNet grant that funded the PlantID section of this paper

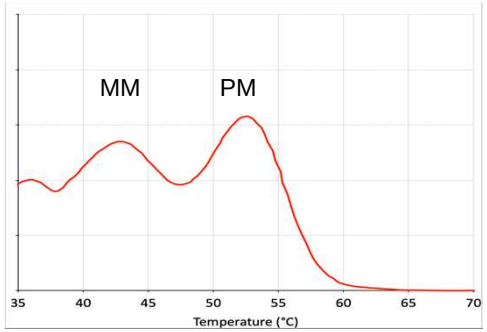
Anthony J. Brookes (AJB)

Oversight and leadership of the activities of SG, NZ, PF, OL and CV. Majority input in to the development of decision tree diagram in Figure 6, technical advice on the development of the core ArrayDASH technology, extensive assistance in the data interpretation of the ArrayDASH (beyond simple homozygous calling and basic detection of heterozygous samples) and majority input into the comparison of ArrayDASH results to the NGS data generated by the Spanish team. Also obtained the READNA funding for the bulk of the work.

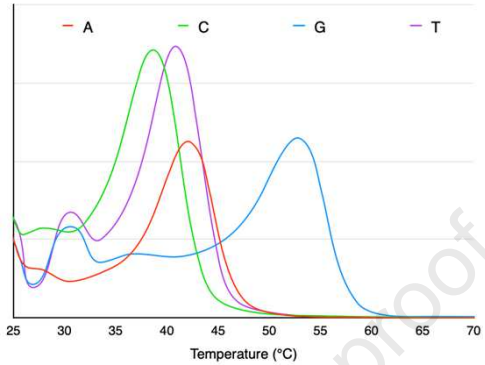
Array Based Dynamic Allele Specific Hybridisaton (Array-DASH):
dynamic microarrays for simultaneous genotyping, fingerprinting, scanning and re-sequencing

Spencer J. Gibson[1], Nathalie Zahra[2], Peter. J. Freeman[3], Caroline Howard[4], Owen Lancaster[5], Colin Veal[1], Maria Casadellà[6], Roger Paredes[6], Marc Noguera-Julian[6], Adrian Slater[7] and Anthony J. Brookes[1]*

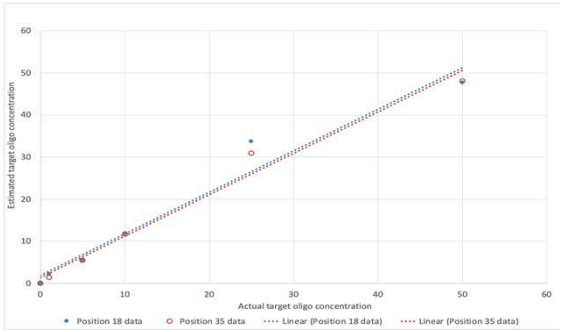
Genreal assay performance



Melt-curve based analysis separates signals from perfectly matched (PM) and mismatched (MM) targets



Re-sequencing accuracy of >99% demonstrated

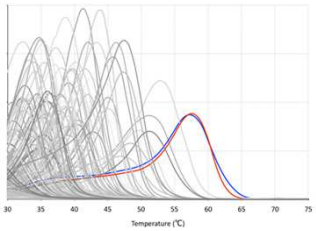


Semi-quantitative detection of PM allele to 1%

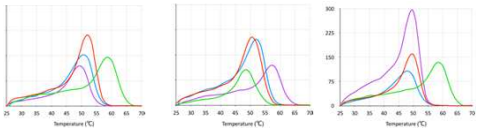
Application to challenging real-world examples

HIV simultaneous genotyping and re-sequencing

Genotyping Probe results

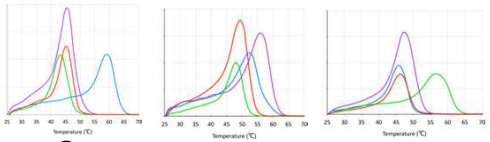


Protease Codon 10 re-sequencing



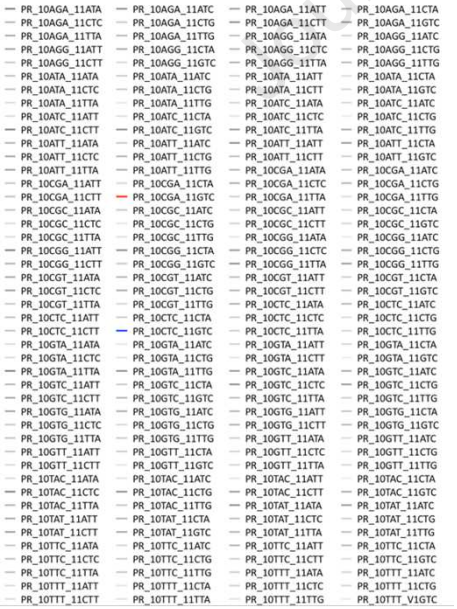
C T C

Protease Codon 11 re-sequencing



G T C

Genotyping Probes on array

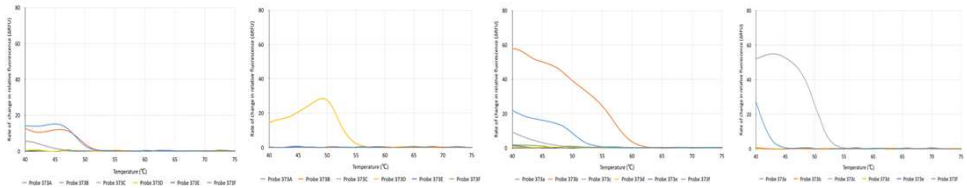


Correct result PR_10CTC_11GTC

Discrimination of *Hypericum* species based on Ribosomal internal transcribed spacer region probes.

Probe set	Reference Sequence.	Probes giving detectable results.			
		H. perforatum	*H. kouchenses*	*H. maculartum*	*H. athoum*
106	AGAAGTGTAAGGCTCCC	105A	105F	105A	105B
116	GCTCCCGGCTGTGCCGG	116A	116C	116F	116A
150	CGGGGGCCT-CCTTCTGT	150A	150C	150B	150B
373	GAGACAATCGGGAATAG	373A + 373B	373D	373B	373C

Probe 373 results



H. perforatum *H. kouchenses* *H. maculartum* *H. athoum*

Array-based Dynamic Allele Specific Hybridization (Array-DASH): optimization-free microarray processing for multiple simultaneous genomic assays.

Spencer J. Gibson[1], Nathalie Zahra[2], Peter. J. Freeman[3], Caroline Howard[4], Owen Lancaster[1], Colin Veal[1], Maria Casadellà Fontdevila[5], Roger Paredes[5], Marc Noguera-Julian[5], Adrian Slater[6] and Anthony J. Brookes[1]*

[1] Department of Genetics, Adrian Building, University of Leicester, University Road, Leicester, Leicestershire, LE17RH, UK.

[2] Anglia Ruskin University, Department of Biomedical and Forensic Science, SCI004, East Road, Cambridge CB1 1PT UK.

[3] Division of Informatics, Imaging & Data Science, Faculty of Biology, Medicine and Health, The University of Manchester, G.725, Stopford Building, Oxford Road, Manchester, M13 9PT

[4] Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

[5] Institut de Recerca de la SIDA – IrsiCaixa, Hospital Universitari Germans Trias i Pujol, Crta. de Canyet s/n, Planta 2a, 08916 Badalona, Catalonia, Spain

[6] Biomolecular Technology Group, School of Allied Health Sciences, De Montfort University, Hawthorn Building HB1.12, The Gateway, Leicester, Leicestershire, LE19BH, UK.(Honorary)

* Brookes A.J.. Tel: +44 116 252 3401; Email: ajb97@leicester.ac.uk

ABSTRACT

We report proof-of-principle experiments regarding a dynamic microarray protocol enabling accurate and semi-quantitative DNA analysis for re-sequencing, fingerprinting and genotyping. Single-stranded target molecules hybridise to surface-bound probes during initial gradual cooling with high-fidelity. Real-time tracking of target denaturation (via fluorescence) during a 'dynamic' gradual heating phase permits 'melt-curve' analysis. The probe most closely matching the target sequence is identified based on the highest melting temperature. We demonstrated a >99% re-sequencing accuracy and a potential detection rate of 1% for SNPs. Experiments employing *Hypericum* ribosomal ITS regions and HIV genomes illustrated a reliable detection level of 5% plus simultaneous re-sequencing and genotyping. Such performance suggests a range of potential real-world applications involving rapid sequence interrogation, for example, in the Covid-19 pandemic. Guidance is offered towards the development of a commercial platform and dedicated software required to bring this technique into mainstream science.

1. INTRODUCTION

Hybridisation of test sequences to oligonucleotide microarrays is a widely used method for high-throughput analysis of both RNA and DNA. Researchers have reported various applications, with predominant areas of use being quantitative gene expression analysis and comparative genome hybridisation. Approaches using traditional microarrays for screening subtle genetic variants, such as SNPs, have required careful control of reaction conditions and probe design. This is primarily because some sets of highly similar oligonucleotide probes used to interrogate particular variants may not distinguish the cognate alleles reliably, under the specific 'static' conditions used (i.e., the annealing and washing stringencies applied to the array).

Many teams have optimised microarray designs and reaction conditions for similar sequence discernment. For example, Suzuki *et al.*[1] showed that a probe length of between 19-21bp was most effective in discriminating single base pair differences. Other reports[2,3] have characterised the effect of temperature and other reaction conditions (such as buffer type, salt concentration and the temperature of the stringency wash) on the stability of perfectly matched probe:target hybrids compared to those containing one or more mismatches. Horiuchi[4] and his team evaluated the method's performance and thereby showed that the rate of false positive SNP assignments in the rice genome was as high as 23% using an Affymetrix GeneChip. However, others have used the Affymetrix p53 GeneChip on bladder cancer samples and have been able to tune the method to minimise the false positive rate down to 14% but with the negative consequence that only 84% of extant mutations were detected[5]. Then again, a similar level of "true SNP" detection (determined by Sanger sequencing with radiolabelled nucleotides and manual interpretation of the same samples) when compared to performance using the same chip has been reported[6] for primary lung cancer (identifying 81% of p53 gene mutations in 100 samples), but in this case the false positive rate was reduced to 2%. Regarding sensitivity, Wikman *et al.*, demonstrated the ability to detect as little as 1% mutant sequence in a background of 99% wild type in bladder tumour DNA using the same p53 chip[5].

The take-home message from these and many other such studies is that it is challenging to apply traditional microarrays for DNA sequence analysis and achieve good all-round performance using standardised assay conditions. Instead, the method requires optimisation for each specific set of targets, to enhance whatever performance characteristics are most critical to that use case.

We first showed a different strategy for array hybridisation in 1999 when we described the DASH (dynamic allele specific hybridisation) approach[7]. The key innovation in DASH was the continuous monitoring of denaturation events between the probe and the target, induced by 'dynamically' increasing temperature of the reaction following target hybridisation. It represented the data obtained as real-time melt-curve graphs.

Previous incarnations of the DASH method have used different approaches to fluorescence signal

Array-based Dynamic Allele Specific Hybridization (Array-DASH): optimization-free microarray processing for multiple simultaneous genomic assays.

Spencer J. Gibson[1], Nathalie Zahra[2], Peter. J. Freeman[3], Caroline Howard[4], Owen Lancaster[1], Colin Veal[1], Maria Casadellà Fontdevila[5], Roger Paredes[5], Marc Noguera-Julian[5], Adrian Slater[6] and Anthony J. Brookes[1]*

[1] Department of Genetics, Adrian Building, University of Leicester, University Road, Leicester, Leicestershire, LE17RH, UK.

[2] Anglia Ruskin University, Department of Biomedical and Forensic Science, SCI004, East Road, Cambridge CB1 1PT UK.

[3] Division of Informatics, Imaging & Data Science, Faculty of Biology, Medicine and Health, The University of Manchester, G.725, Stopford Building, Oxford Road, Manchester, M13 9PT

[4] Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

[5] Institut de Recerca de la SIDA – IrsiCaixa, Hospital Universitari Germans Trias i Pujol, Crta. de Canyet s/n, Planta 2a, 08916 Badalona, Catalonia, Spain

[6] Biomolecular Technology Group, School of Allied Health Sciences, De Montfort University, Hawthorn Building HB1.12, The Gateway, Leicester, Leicestershire, LE19BH, UK.(Honorary)

* Brookes A.J.. Tel: +44 116 252 3401; Email: ajb97@leicester.ac.uk

ABSTRACT

We report proof-of-principle experiments regarding a dynamic microarray protocol enabling accurate and semi-quantitative DNA analysis for re-sequencing, fingerprinting and genotyping. Single-stranded target molecules hybridise to surface-bound probes during initial gradual cooling with high-fidelity. Real-time tracking of target denaturation (via fluorescence) during a 'dynamic' gradual heating phase permits 'melt-curve' analysis. The probe most closely matching the target sequence is identified based on the highest melting temperature. We demonstrated a >99% re-sequencing accuracy and a potential detection rate of 1% for SNPs. Experiments employing *Hypericum* ribosomal ITS regions and HIV genomes illustrated a reliable detection level of 5% plus simultaneous re-sequencing and genotyping. Such performance suggests a range of potential real-world applications involving rapid sequence interrogation, for example, in the Covid-19 pandemic. Guidance is offered towards the development of a commercial platform and dedicated software required to bring this technique into mainstream science.

1. INTRODUCTION

Hybridisation of test sequences to oligonucleotide microarrays is a widely used method for high-throughput analysis of both RNA and DNA. Researchers have reported various applications, with predominant areas of use being quantitative gene expression analysis and comparative genome hybridisation. Approaches using traditional microarrays for screening subtle genetic variants, such as SNPs, have required careful control of reaction conditions and probe design. This is primarily because some sets of highly similar oligonucleotide probes used to interrogate particular variants may not distinguish the cognate alleles reliably, under the specific 'static' conditions used (i.e., the annealing and washing stringencies applied to the array).

Many teams have optimised microarray designs and reaction conditions for similar sequence discernment. For example, Suzuki *et al.*[1] showed that a probe length of between 19-21bp was most effective in discriminating single base pair differences. Other reports[2,3] have characterised the effect of temperature and other reaction conditions (such as buffer type, salt concentration and the temperature of the stringency wash) on the stability of perfectly matched probe:target hybrids compared to those containing one or more mismatches. Horiuchi[4] and his team evaluated the method's performance and thereby showed that the rate of false positive SNP assignments in the rice genome was as high as 23% using an Affymetrix GeneChip. However, others have used the Affymetrix p53 GeneChip on bladder cancer samples and have been able to tune the method to minimise the false positive rate down to 14% but with the negative consequence that only 84% of extant mutations were detected[5]. Then again, a similar level of "true SNP" detection (determined by Sanger sequencing with radiolabelled nucleotides and manual interpretation of the same samples) when compared to performance using the same chip has been reported[6] for primary lung cancer (identifying 81% of p53 gene mutations in 100 samples), but in this case the false positive rate was reduced to 2%. Regarding sensitivity, Wikman *et al.*, demonstrated the ability to detect as little as 1% mutant sequence in a background of 99% wild type in bladder tumour DNA using the same p53 chip[5].

The take-home message from these and many other such studies is that it is challenging to apply traditional microarrays for DNA sequence analysis and achieve good all-round performance using standardised assay conditions. Instead, the method requires optimisation for each specific set of targets, to enhance whatever performance characteristics are most critical to that use case.

We first showed a different strategy for array hybridisation in 1999 when we described the DASH (dynamic allele specific hybridisation) approach[7]. The key innovation in DASH was the continuous monitoring of denaturation events between the probe and the target, induced by 'dynamically' increasing temperature of the reaction following target hybridisation. It represented the data obtained as real-time melt-curve graphs.

Previous incarnations of the DASH method have used different approaches to fluorescence signal

This highlighted section is the figure. Please either typeset as is, or treat as fixed graphic

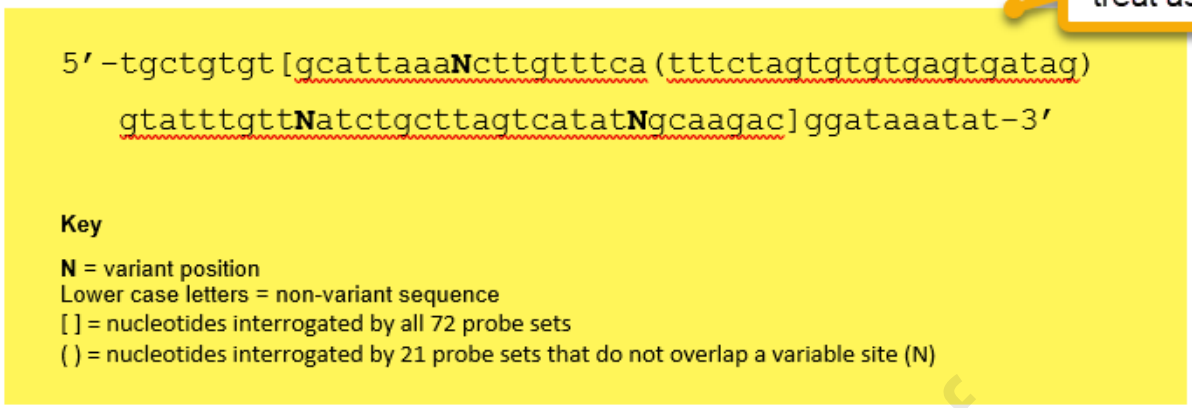


Figure 3: Diagram showing the positions in LSCAN control oligonucleotides where there are one or two mismatches to the microarray probes.

The synthetic LSCAN target oligonucleotide sequence is shown as synthesised. This is the reverse compliment of the CACNA2D1 gene sequence encoding the calcium voltage-gated channel auxiliary subunit alpha2delta 1 (chromosome 7 bases 82376203-82376292, human genome build GRCh38.p13) reference sequence used to design the probes on the microarray.

5′-tgctgtgt[gcattaaa**N**cttgtttca(tttctagtgtgtgagtgatag)

gtatttgtt**N**atctgcttagtcatat**N**gcaagac]ggataaatat-3′

Key

N = variant position
Lower case letters = non-variant sequence
[ ] = nucleotides interrogated by all 72 probe sets
( ) = nucleotides interrogated by 21 probe sets that do not overlap a variable site (N)

Key Phrases

- Sensitive melt-curve microarray
- Re-sequencing & genotyping
- HIV & Hypericum typing