HHS Public Access

Author manuscript

Nat Chem. Author manuscript; available in PMC 2018 January 29.

Published in final edited form as:

Nat Chem. 2017 December ; 9(12): 1222–1228. doi:10.1038/nchem.2820.

Modular probes for enriching and detecting complex nucleic acid sequences

Juexiao Sherry Wang[1,2,†], **Yan Helen Yan**[1,2,†], and **David Yu Zhang**[1,2,*]

[1]Systems, Synthetic, and Physical Biology, Rice University, Houston, Texas 77030, USA.

[2]Department of Bioengineering, Rice University, Houston, Texas 77030, USA.

Abstract

Complex DNA sequences are difficult to detect and profile, but are important contributors to human health and disease. Existing hybridization probes lack the capability to selectively bind and enrich hypervariable, long or repetitive sequences. Here, we present a generalized strategy for constructing modular hybridization probes (M-Probes) that overcomes these challenges. We demonstrate that M-Probes can tolerate sequence variations of up to 7 nt at prescribed positions while maintaining single nucleotide sensitivity at other positions. M-Probes are also shown to be capable of sequence-selectively binding a continuous DNA sequence of more than 500 nt. Furthermore, we show that M-Probes can detect genes with triplet repeats exceeding a programmed threshold. As a demonstration of this technology, we have developed a hybrid capture method to determine the exact triplet repeat expansion number in the Huntington's gene of genomic DNA using quantitative PCR.

The human genome comprises over 3 billion nucleotides, roughly 96% of which[1–3] can now be routinely analysed by next-generation sequencing (NGS) platforms[4,5]. The other 4% is composed of DNA sequences that are long, repetitive and/or hypervariable, and which are therefore challenging to detect and profile by NGS[6,7] or other DNA analysis platforms[8,9]. Simultaneously, these complex regions have important effects on human health and disease. Examples include microsatellite expansion in cancer[10], triplet expansion in Huntington's and other neurological diseases[11,12], VDJ recombination in leukocytes[13,14] and structural variants in cancer[15–17].

Hybridization probes have been long used to capture and enrich specific DNA sequences to ease downstream analysis[8,18–20] and have recently gained popularity for targeted sequencing

*Correspondence and requests for materials should be addressed to D.Y.Z., dyz1@rice.edu.
†These authors contributed equally to this work.

Author contributions

J.S.W., Y.H.Y. and D.Y.Z. conceived the project. J.S.W. and Y.H.Y. performed experiments and data analysis. J.S.W., Y.H.Y. and D.Y.Z. wrote the manuscript. J.S.W. and Y.H.Y. contributed equally to this work.

Additional information

Supplementary information is available in the online version of the paper.

Competing financial interests

There is one patent pending on M-Probes described in this work.

of panels of disease-specific genes[21] or the whole exome[22]. However, existing hybridization probes face three limitations that render them unsuitable for binding and enriching complex nucleic acid sequences. First, the length of the probe is typically limited to between 20 and 50 nt to maintain sequence selectivity. Second, the target sequence in general cannot be highly repetitive. Third, probes are unable to differentiate between important and incidental variations in the target sequence.

Here, we present modular hybridization probes (M-Probes) that overcome the above limitations, enabling selective binding of long, complex and repetitive nucleic acid sequences from genomic DNA. The M-Probe is constructed from multiple probe segments, each targeting one section of a potentially long target sequence. The modular nature of the segments circumvents the synthesis limitations of oligonucleotides. Simultaneously, a competitive hybridization reaction is rationally designed to ensure single nucleotide specificity despite probe and target lengths of over 500 nt. The junctions between the probe segments tolerate up to 7 nt of sequence variation without significant effect on binding affinity, while even 1 nt variants at other locations result in more than threefold reduction. Finally, because individual segments can be formulated separately and then combined, the chance of probe malformation due to misaligned binding for repetitive sequences is minimized. We have used M-Probes to directly detect high-concentration target sequences (using a fluorophore-functionalized M-Probe) and to enrich low-concentration target sequences from heterogeneous genomic DNA samples (using magnetic beads and a biotin-functionalized M-Probe).

Results

M-Probe design and validation

Figure 1a shows the general structure and construction of an M-Probe for direct detection of a target nucleic acid sequence. An M-Probe consists of a left universal segment *u*, *n* internal segments labelled s_1 to s_n and a right termination segment *t*. In the termination segment *t*, the upper oligo is shorter than the lower oligo by a number of nucleotides, and the single-stranded nucleotides on the rightmost lower strand are referred to as the ‘toehold’[23].

Each segment consists of two oligonucleotides hybridized to each other via a horizontal region; in the *s* and *t* segments, these horizontal regions’ sequences are target-specific. Throughout this Article, the lower oligonucleotides have sequence complementary to subsequences of the target, and the upper oligonucleotides have sequence identical to subsequences of the target. Different segments are hybridized to each other via two vertical ‘arms’ with sequences independent of the target. For efficient formulation, the two arms each have a unique sequence that is *in silico* designed to be orthogonal to the other, and also unlikely to bind to the human genome.

Following the hybridization reaction with the target sequence, the upper M-Probe oligos are released as a multistranded complex (Fig. 1b). Afterwards, the released multistranded complex can re-associate with the product and induce the reverse reaction. This reaction reversibility allows the hybridization between the target and M-Probe to be selective, regardless of the length of the target, following principles described in ref. 24 (see

Supplementary Section 1 for design details). Figure 1c shows the fluorescence response of a sequence-selective $n = 1$ M-Probe to its target, a 43 nt sequence, as well as two single-nucleotide variants of the target. Simultaneously, the M-Probe is not poisoned by long-lived reaction intermediates with the variants, and the addition of target at 2 h results in an immediate and strong fluorescence response.

Programmed sequence variation tolerance

One technical challenge for many hybridization-based enrichment and detection methods is to tolerate potential single-nucleotide polymorphisms (SNPs) at known locations. Inherited SNPs are frequent in the human genome, with the literature reporting SNP frequencies of roughly 1 per 1,000 nt on average[25]. Many SNPs are intronic or synonymous mutations with no effect on protein sequence, but they may interfere with hybridization probe detection or enrichment due to their close proximity to clinically or scientifically important sequence variations. As one example, rs1050171 is a synonymous SNP in the EGFR gene (c. 2361G>A) with a 43% allele frequency in the human population. It is eight nucleotides away from the c.2369C>T (T790M) mutation that confers resistance to the cancer drug erlotinib. The 1000 Genomes and dbSNP databases[3,26] provide sequence, position and frequency information for SNPs with allele frequencies of 0.5% or higher in the human genome.

The M-Probe uniquely offers tolerance to sequence variations at the segment junctions (Fig. 2a), and sequence changes, insertions and deletions at these positions have only a small to insignificant effect on the overall hybridization reaction standard free energy. In contrast, target sequence variations at positions that hybridize to the segments result in bulges or mismatch bubbles that destabilize the hybridization product and render $\Delta G^{\circ}_{\mathrm{Hyb}}$ significantly more positive as compared to the intended target sequence, resulting in lower hybridization yield and fluorescence. Experimentally, we observe that up to 7 nt variations at the segment junctions have little impact on M-Probe hybridization yield (Fig. 2b,d), but even single nucleotide variants in segments s_1 and t result in a severe reduction in binding yield (Fig. 2c,d).For targets with known potential variations at particular loci, the M-Probe can be designed so that these loci correspond to the segment junctions. There is significant sequence variability in terms of number of nucleotides of insertion tolerated due to varying secondary structures. We believe, based on our experiments and published thermodynamic parameters, that a large majority of 7 nt insertions should be tolerated. However, there is no hard maximum on the number of insertion nucleotides tolerated; for example, a favourable target sequence bearing a 100 nt insertion that forms a perfect hairpin would probably still be tolerated.

Combinatorial M-Probe formation

Another feature of the modular construction of the M-Probe is that multiple different internal segments can be combinatorially combined to generate many different M-Probes for different target sequences (Fig. 3a). Given m_i instances for each segment s_i, the total number of M-Probes that can be constructed is given by

$$\text{Number of unique M- Probes} = m_t \prod_{i=1}^{n} m_i$$

where m_t is the number of instances of the terminal segment t. The number of oligonucleotides used to construct these, in contrast, scales with the sum of m_i:

$$\text{Number of oligonucleotides} = 2\left(1 + m_t + \sum_{i=1}^{n} m_i\right)$$

For large n and m_i values, combinatorial formulation significantly reduces the number of oligonucleotides needed to detect or enrich sequences. In human T cells, the TCR-β gene undergoes VDJ recombination in which 1 V, 1 D and 1 J gene region are selected from 48 V, 2 D and 13 J genes segments, respectively (Fig. 3b). Random deletions and non-templated insertions occur at the V–D and D–J junctions to provide further T-cell receptor diversity to facilitate recognition of diverse antigens. Combinatorially formulated M-Probes that tolerate sequence variation at the VDJ junction are thus well suited for hybridization-based detection and enrichment of the recombined TCR-β gene.

Because of the short length and high sequence variability in the D gene region, we elected to consider the entire D region as variable and designed the M-Probes to be $n = 1$, with the s_1 segment corresponding to the V region and the t segment corresponding to the J region. $m_1 = 8$ and $m_t = 6$ different instances of the s_1 and t segments were designed, allowing the detection of 48 combinatorially recombined VDJ sequences (Fig. 3c). The bulge formed between the segments upon binding an M-Probe to its intended target varies in length between 8 and 32 nt. See Supplementary Sections 2 and 3 for details of the M-Probe design. Figure 3d presents a summary of the hybridization between 48 TCR-β sequence targets and the 48 M-Probes. The main diagonal corresponds to endpoint fluorescence of the 48 on-target hybridization reactions in which the target perfectly matches the M-Probe in the identity of the s_1 and t regions (see Supplementary Section 4 for data acquisition details). The dark blue off-diagonal squares correspond to 576 off-target hybridization reactions in which the target matches the M-Probe in the identity of either s_1 or t, but not both. Grey squares denote combinations in which both the s_1 and t segments of the M-Probe are mismatched to the target; these were not tested as they were judged to be unlikely to yield a significant hybridization response.

Of the 624 hybridization reactions experimentally characterized, all off-target hybridization experiments generated less than 0.6 a.u. of fluorescence, while 43 (90%) of the on-target hybridization generated more than 0.6 a.u. of fluorescence and 30 (63%) generated more than 1.2 a.u. of fluorescence (Fig. 3e). Thus, the predominant failure mode appears to be that a fraction of the M-Probes fail to hybridize significantly with their matched target sequences. Possible reasons for underperformance include unfavourable hybridization thermodynamics due to inaccurate estimates of the destabilizing effect of large bulges and slow hybridization kinetics due to secondary structure in the target sequence. Empirical optimization of M-Probe sequence design may overcome thermodynamics errors, and the operation of M-Probes at higher temperatures may accelerate hybridization kinetics.

Long targets

M-Probes with $n \geq 1$ have multiple target-specific segments (including t) and can circumvent oligonucleotide synthesis limitations to probe longer continuous target sequences. For example, given an oligonucleotide synthesis limitation of L nucleotides (L = 100 for a standard oligo, L = 200 for an IDT Ultramer oligo (long synthetic DNA synthesized via IDT proprietary methods)), each of the n internal s segments can probe ($L - 2A$) nucleotides (where A is the length of the arm sequence), and the terminal t segment can probe ($L - A$) nucleotides. An n internal segment M-Probe can thus probe a maximum length L_M of $L_M = (n + 1)\ L - (2n + 1)\ A$ continuous nucleotides. From the above equation it is clear that the M-Probe benefits from shorter arm lengths, A. The minimum length of A for stable formation of the M-Probe depends on arm sequence, temperature and buffer salinity. At 37–45 °C and 1× PBS, A = 22 is sufficient for stability for most arm sequences. For L = 180 and A = 22, an n = 2 M-Probe can probe up to 430 nt and an n = 3 M-Probe can probe up to 564 nt.

M-Probes retain their high sequence selectivity even when binding long DNA targets. Figure 4 shows the detection of two SNPs within two different 560 nt targeting regions. The targets for these experiments are amplicons from the NA18562 and NA18537 cell line genomic DNA, which differ by only a single nucleotide in the middle of the M-Probe targeting region. There has been no previous demonstrations of single-nucleotide selectivity in DNA hybridization probes for probe lengths longer than ~50 nt. Consequently, M-Probes increased the effective length range of allele-specific detection and enrichment by more than tenfold and could potentially be used as a novel method for confirming sequence. We also designed and tested M-Probes with targeting sequences that were 99, 160, 218 and 430 nt long and obtained the expected results (see Supplementary Section 5 for details).

Trinucleotide repeat profiling

DNA trinucleotide repeat expansion is a difficult problem for standard molecular analysis technologies due to the long lengths and repetitive sequences involved. There are over 20 hereditary neurological disorders linked to an increased number of DNA trinucleotide repeats[12]; examples include Huntington's disease (CAG repeats), fragile X syndrome (CGG repeats) and Friedreich's ataxis (GAA repeats). In each of these diseases, the relevant genes of affected patients feature a triplet repeat count number above a threshold amount (for example, 27 repeats for Huntington's disease), as compared to healthy individuals' genes, which have a sub-threshold number of triplet repeats.

Figure 5a presents a schematic of M-Probes designed for profiling CAG triplet repeats in the Huntington's gene, HTT. A non-repetitive HTT-specific sequence in the t segment of each M-Probe ensures that M-Probe binding is specific to the HTT gene and not other genomic regions bearing CAG triplet repeats. Although various M-Probe formulation protocols generally produce similar formation yields (Supplementary Fig. 1–5), for the repeat sequence M-Probes in this section, segments were individually annealed and subsequently combined (See Supplementary Methods). A series of M-Probes were designed, each targeting a different number of CAG repeats, and any HTT gene sequence with a repeat number equal to or exceeding the M-Probe's repeat number will elicit a positive signal.

Thus, when aliquots of an HTT gene or amplicon are reacted with the series of M-Probes, the longest M-Probe that still generates a positive signal indicates the number of triplet repeats.

Figure 5b shows the fluorescence response of a conditionally fluorescent M-Probe targeting 18 CAG repeats to five different DNA oligonucleotide targets bearing 12, 15, 18, 21 and 24 repeats (labelled T12, T15, T18, T21 and T24, respectively). The M-Probe shows significant binding to T18, T21 and T24, but not to T12 and T15, confirming that this M-Probe functions as designed in acting as a programmable high-pass filter on trinucleotide repeat number. Figure 5c shows a summary of the response between different M-Probes and synthetic oligonucleotide targets, and significant hybridization is observed only when the number of target repeats equals or exceeds the M-Probe repeat number. Similar M-Probes were also verified to profile the triplet repeat number for CGG repeats (associated with fragile X syndrome) and GAA repeats (associated with Friedriech's ataxia). (See Supplementary Section 6 and Supplementary Fig. 6–4 for details and results.)

To apply M-Probes to profiling the triplet repeat number in HTT in genomic DNA samples, biotin-functionalized M-Probes were used to selectively bind DNA with HTT sequence exceeding the threshold number of triplet repeats (Fig. 5d). The genomic DNA sample was first pre-amplified with a five-cycle PCR protocol to generate amplicons bearing the HTT triplet repeats as well as a minimal 20 nt upstream and 14 nt downstream from the repeats. Amplicons generated in this fashion do not have long 5′ and 3′ overhangs that may interfere with hybridization to M-Probes (due to secondary structure, and so on). These amplicons were subsequently incubated with the appropriate M-Probe and captured by streptavidin-coated magnetic beads. Unbound DNA molecules were removed using a wash step. The captured DNA was eluted and quantitated using qPCR (Fig. 5e).

Amplification of HTT genes with less than the threshold repeat number (number of triplets in the M-Probe) showed a significantly higher cycle threshold (Ct) than the HTT genes exceeding the threshold repeat number. By designing two different M-Probes, one targeting 9 repeats and one targeting 27 repeats, we could control for sample variability and determine the potential disease status through the difference in the observed Ct values (ΔCt). Small (<2) ΔCt values indicate that at least one of the two HTT gene copies exceeds 27 repeats, and large (>5) ΔCt values indicate the opposite. Residual amplification of the low-repeat-number HTT genes is probably due to non-specific binding of genomic DNA to the magnetic beads (data not shown).

Figure 5f summarizes the observed results for seven genomic DNA samples, five with known HTT genotypes and two unknown. Our method correctly identified the length status of the five known samples and determined that the NA18537 and NA18524 samples both only possessed HTT genes with fewer than 27 CAG repeats. The two M-Probe systems (here targeting 9 and 27 repeats) represent the minimal protocol needed for determining disease likelihood in an unknown genomic DNA sample.

More precise quantitation of the HTT triplet repeat number can be achieved by extending the method to include more M-Probes with varying triplet repeat thresholds. To demonstrate this

point, we designed five different M-Probes targeting 33, 35, 36, 37 and 39 CAG repeats, and applied it to the NA20248 genomic DNA sample. The experimental Ct values for the M-Probes targeting 37 and 39 repeats were more than five cycles higher than for M-Probes targeting 33, 35 and 36 repeats, suggesting correctly that the sample had one HTT gene copy with exactly 36 CAG repeats (Fig. 5g).

In addition to the hybrid-capture workflow we present here, an alternative approach to profiling triplet repeats using M-Probes is to amplify the HTT gene to above nanomolar concentrations and then directly react the amplicons with conditionally fluorescent M-Probes. The relative advantage of this second approach is that the solid-phase separation steps are avoided, reducing the total hands-on time. Its relative disadvantage is that open-tube steps on high-concentration amplicons are likely to lead to laboratory contamination and are therefore undesirable in diagnostic settings. Proof-of-concept experiments using the secondary approach to profile HTT triplet repeat number are provided in Supplementary Section 6. Both approaches can reliably detect repeat expansion with single repeat resolution in a small range of expansion (for example, 27–40 for Huntington's disease)—this is difficult to achieve using previously reported methods[27–30]. A larger range of expansion can also be profiled by using M-Probes with more and/or longer segments.

Discussion

Conceptually, the M-Probe can be thought of as a multistranded equivalent of the toehold probe[24,31], in which the probe (the lower oligo) and the target-mimic (the upper oligo) sequences are distributed across multiple oligonucleotides connected by arms. The main advantages of the M-Probe over toehold probes are as follows: (1) it enables robustness to target sequence variation at/near the segment boundaries; (2) it allows longer target regions because limitations on oligonucleotide synthesis lengths are independent of the number of segments n; and (3) it decouples the expensive functionalized oligonucleotides from non-functionalized target-specific oligonucleotides, thereby reducing prototyping costs. The X-Probe, which we have presented previously[32], can be thought of as a special case of the M-Probe with $n = 0$, which captures benefit (3) but not (1) or (2). For short target sequences, the performances of conditionally fluorescent toehold probes, X-Probes and M-Probes are similar (Supplementary Fig. 1–5).

As far as we are aware, the M-Probe represents the first experimentally demonstrated probe where single-nucleotide selectivity can be achieved simultaneously with tolerance to multi-nucleotide variation at specified positions. Previously, researchers have incorporated degenerate nucleotide mixtures (for example, N[18,33]) or universal artificial nucleotides (for example, inosine[34]) in probes to confer sequence variation tolerance, but such approaches do not equally tolerate insertions, deletions and replacements. Furthermore, these approaches are generally not compatible with double-stranded probes, which allow single-nucleotide selectivity across long target regions.

Regarding target length, we have demonstrated an M-Probe that sequence-selectively binds a 560-nt-long target sequence from human genomic DNA. This is the longest continuous target sequence sequence-selectively hybridized by more than a factor of 10 (refs 20, 31).

Advances in oligonucleotide synthesis and purification technologies will help realize the full potential of M-Probes in sequence-selective detection and enrichment of long nucleic acid sequences. Extension of the M-Probe approach to significantly longer target regions may enable the specific enrichment of structural variants in cancer genes.

Our proof-of-concept experiments used the M-Probes in two ways: as conditionally fluorescent probes for profiling of PCR amplicons and as biotin-labelled probes for solid-phase capture and enrichment of genomic DNA for subsequent qPCR analysis. The latter method should be easily adaptable to highly multiplexed enrichment for downstream next-generation sequencing analysis, allowing multiplexed targeted profiling of genetic regions with complex variations. Of particular benefit may be enrichment of DNA structural variants (for example, translocations and fusions), RNA alternative splice patterns and other sequences currently difficult to assay with short-read sequencing.

Data availability

All the data presented and analysed in this study are either included within this Article or within the Supplementary Information. Relevant raw data are available from the corresponding author upon reasonable request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank A. Pinto for discussions. This work was funded by the Cancer Prevention Research Institute of Texas (grant RP140132 to D.Y.Z.) and by the National Human Genome Research Institute (grant R01HG008752 to D.Y.Z.).

References

1. Venter JC, et al. The sequence of the human genome. Science. 2001; 291:1304–1351. [PubMed: 11181995]
2. 1000 Genomes Project Consortium. tA map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
3. 1000 Genomes Project Consortium. tA global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
4. Mardis ER. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. 2013; 6:287–303.
5. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30:418–426. [PubMed: 25108476]
6. Quail MA, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341. [PubMed: 22827831]
7. Kircher M, Kelso J. High throughput DNA sequencing concepts and limitations. Bioessays. 2010; 32:524–536. [PubMed: 20486139]
8. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. 2005; 37:549–554. [PubMed: 15838508]
9. Higuchi R, Fockler C, Dollinger G, Watson R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Biotechnology. 1993; 11:1026–1030. [PubMed: 7764001]

10. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. 2013; 155:858–868. [PubMed: 24209623]
11. Duyao M, et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat. Genet. 1993; 4:387–392. [PubMed: 8401587]
12. Iyer RR, Pluciennik A, Napierala M, Wells RD. DNA triplet repeat expansion and mismatch repair. Annu. Rev. Biochem. 2015; 84:199–226. [PubMed: 25580529]
13. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009; 19:1817–1824. [PubMed: 19541912]
14. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012; 135:183–191. [PubMed: 22043864]
15. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat. Rev. Genet. 2006; 7:85–97. [PubMed: 16418744]
16. Yang L, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell. 2013; 153:919–929. [PubMed: 23663786]
17. Wang J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods. 2011; 8:652–654. [PubMed: 21666668]
18. Suggs SV, Wallace RB, Hirose T, Kawashima EH, Itakura K. Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta 2-microglobulin. Proc. Natl Acad. Sci. USA. 1981; 78:6613–6617. [PubMed: 6171820]
19. Parkhurst KM, Parkhurst LJ. Kinetic studies by fluorescence resonance energy transfer employing a double-labeled oligonucleotide: hybridization to the oligonucleotide complement and to single-stranded DNA. Biochemistry. 1995; 34:285–292. [PubMed: 7819209]
20. Tyagi S, Kramer FR. Molecular beacons: probes that fluoresce upon hybridization. Nat. Biotechnol. 1996; 14:303–308. [PubMed: 9630890]
21. Lanman RB, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. PLoS ONE. 2015; 10:e0140712. [PubMed: 26474073]
22. Clark MJ, et al. Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. 2011; 29:908–914. [PubMed: 21947028]
23. Zhang DY, Winfree E. Control of DNA strand displacement kinetics using toehold exchange. J. Am. Chem. Soc. 2009; 131:17303–17314. [PubMed: 19894722]
24. Zhang DY, Chen SX, Yin P. Thermodynamic optimization of nucleic acid hybridization specificity. Nat. Chem. 2012; 4:208–214. [PubMed: 22354435]
25. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]
26. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]
27. Budworth H, McMurray CT. Problems and solutions for the analysis of somatic CAG repeat expansion and their relationship to huntington's disease toxicity. Rare Dis. 2016; 4:e1131885. [PubMed: 27141411]
28. Jama M, Millson A, Miller CE, Lyon E. Triplet repeat primed PCR simplifies testing for Huntington disease. J. Mol. Diagn. 2016; 15:255–262.
29. Bonifazi E, et al. Use of RNA fluorescence *in situ* hybridization in the prenatal molecular diagnosis of myotonic dystrophy type I. Clin Chem. 2006; 52:319–322. [PubMed: 16449216]
30. Kern A, Seitz O. Template-directed ligation on repetitive DNA sequences: a chemical method to probe the length of Huntington DNA. Chem. Sci. 2015; 6:724–728. [PubMed: 28706635]
31. Wu LR, et al. Continuously tunable nucleic acid hybridization probes. Nat. Methods. 2015; 12:1191–1196. [PubMed: 26480474]
32. Wang JS, Zhang DY. Simulation-guided DNA probe design for consistently ultraspecific hybridization. Nat. Chem. 2015; 7:545–553. [PubMed: 26100802]

33. Kwok S, Chang SY, Sninsky JJ, Wang A. A guide to the design and use of mismatched and degenerate primers. Genome Res. 1994; 3:S39–S47.

34. Ohtsuka E, Matsuki S, Ikehara M, Takahashi Y, Matsubara K. An alternative approach to deoxyoligonucleotides as hybridization probes by insertion of deoxyinosine at ambiguous codon positions. J. Biol. Chem. 1985; 260:2605–2608. [PubMed: 3838308]

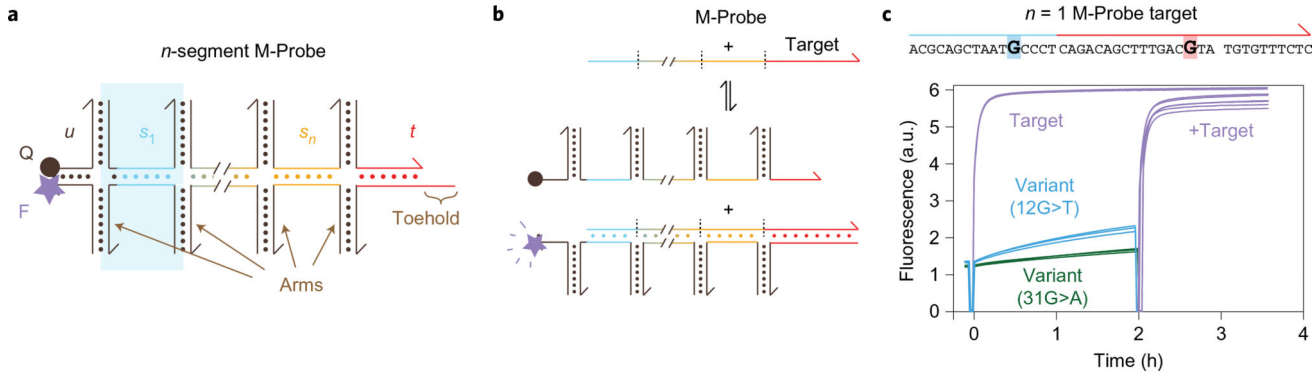

Figure 1. M-Probe design and demonstration

a, A conditionally fluorescent M-Probe bearing *n* internal segments. The lower oligonucleotides have a sequence complementary to sub-sequences of the target, and the upper oligonucleotides have a sequence identical to sub-sequences of the target. The single-stranded nucleotides of the rightmost lower oligo initiate the hybridization reaction and are referred to as the toehold. In the universal segment (*u*) the upper and lower oligos are functionalized with a quencher (Q) and fluorophore (F), respectively. **b**, Hybridization of the M-Probe to the target results in displacement of the upper oligos as a multistranded complex. Fluorescence increases through this process due to delocalization of the fluorophore and quencher. The hybridization reaction is designed to be both reversible and sequence-specific. **c**, Triplicate experimental fluorescence results for response of an $n = 1$ M-Probe (10 nM) to a 43 nt synthetic target oligonucleotide (30 nM) at 37 °C in 1× PBS. Single-nucleotide variants of the target (12G>T and 31G>A, respectively, in the s_1 and *t* segments, highlighted) elicit significantly lower fluorescence signal than the intended target T.

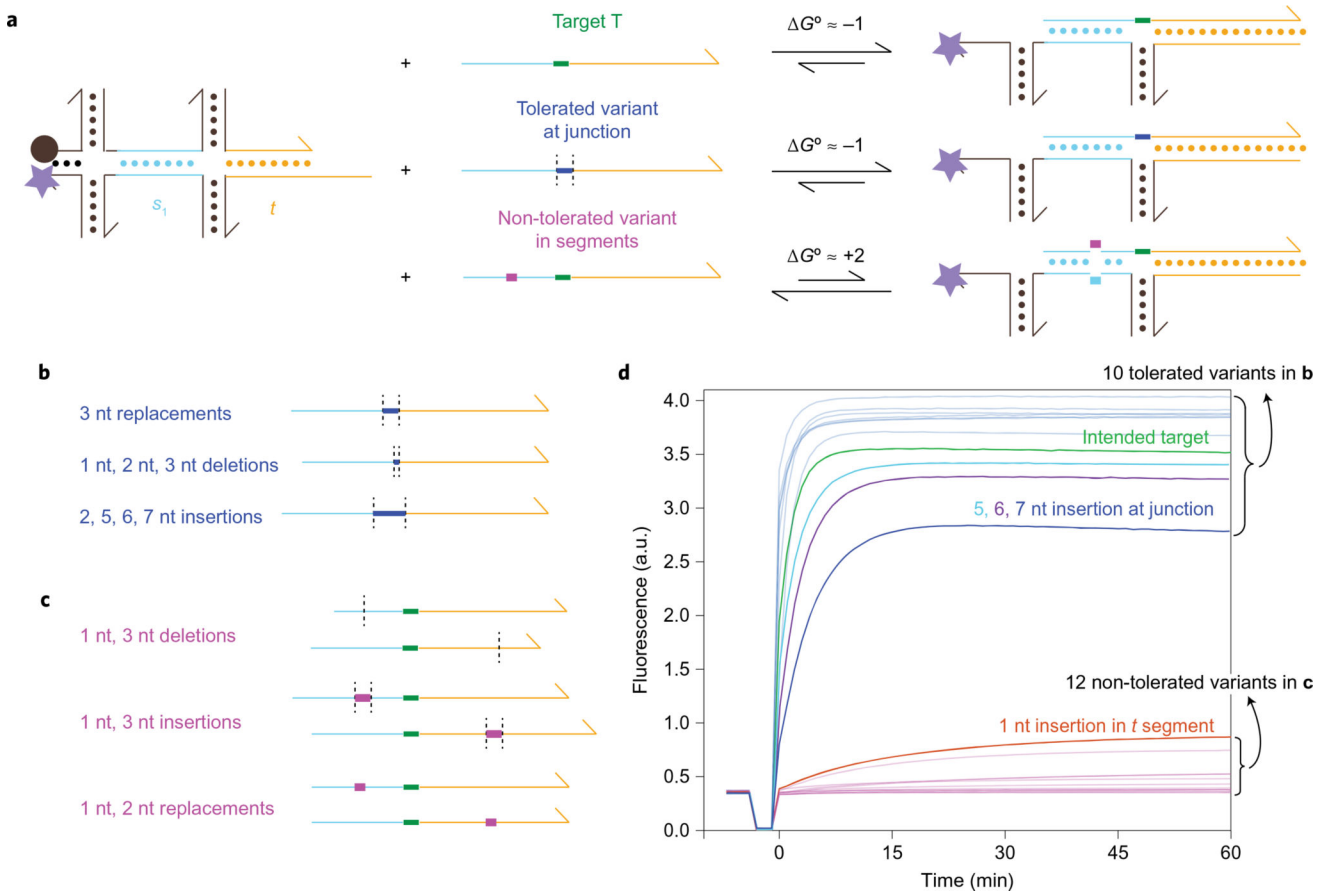

Figure 2. Sequence variation tolerance

a, The M-Probe may be designed such that a number of nucleotides in the intended target are at the segment junction and do not hybridize to any nucleotides in the M-Probe (green region). Sequence variations in this 'tolerant' region have a small effect on the $\Delta G°$ of the hybridization reaction with the M-Probe. In contrast, even single nucleotide variants in other regions lead to large changes in the reaction $\Delta G°$, resulting in significantly lower binding yield. **b**, Ten tolerated variations. **c**, Twelve non-tolerated variations. **d**, Using the same M-Probe, the fluorescence response of targets with tolerated variations of up to 7 nt are not significantly reduced below that of the intended target. Non-tolerated variants, on the other hand, show sharply reduced hybridization yield, even for 1 nt deletion, insertions and replacements.

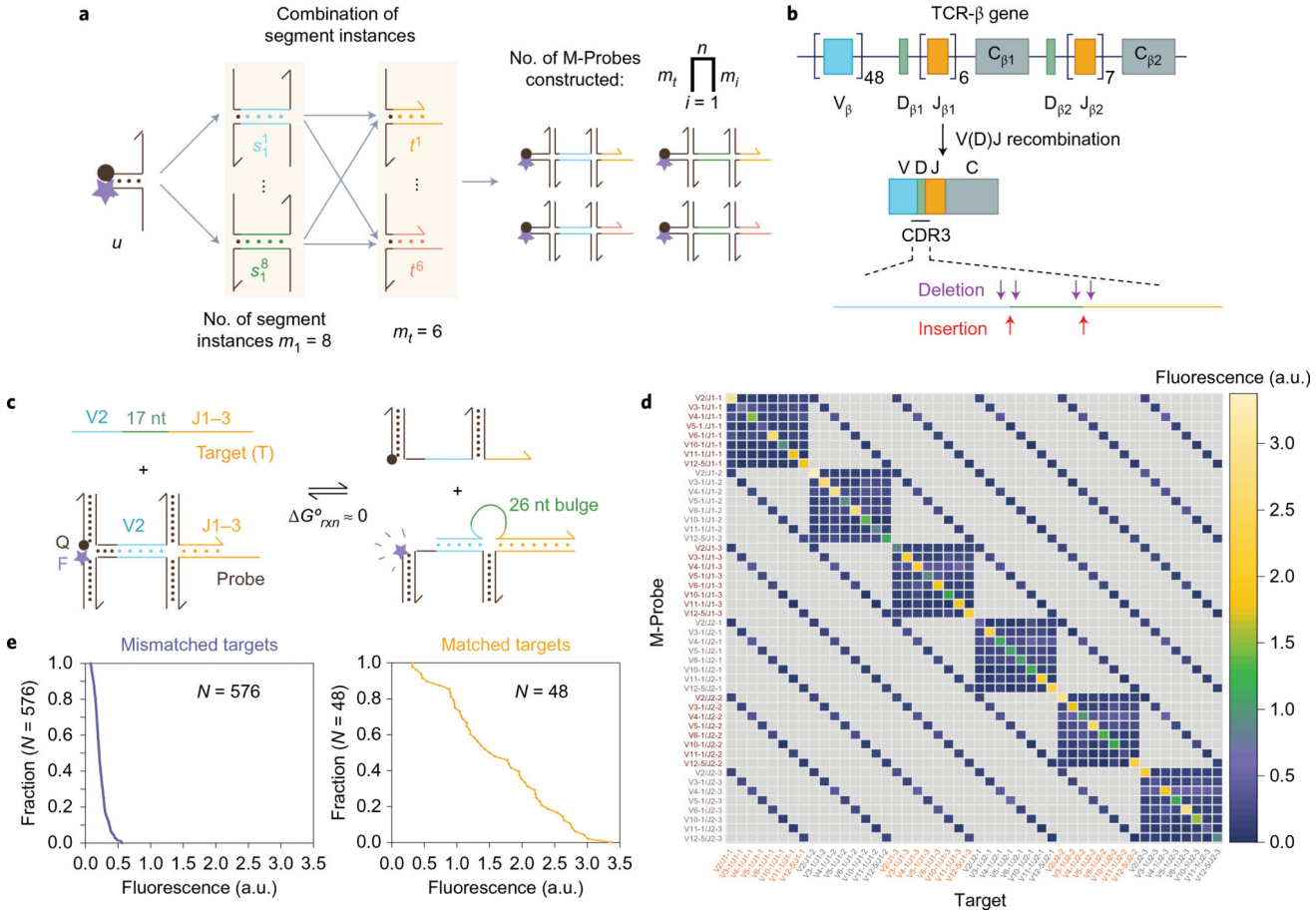

Figure 3. Combinatorial construction of M-Probes for VDJ recombination detection

a, Each target-specific segment can be chosen from a number of modules (m_i). Pairwise combination of modules of adjacent segments allows construction of large number of probes targeting different targets by using limited number of component strands. **b**, Human TCR-β gene VDJ recombination process. Recombination occurs between 48 TRB V (blue), 2 TRB D (green) and 13 TRB J (orange) genes, followed by random deletions and non-templated insertions at the V–D and D–J junctions, resulting in a hypervariable CDR3 region that is important for antigen recognition. **c**, Example hybridization reaction between an M-Probe and a matched target sequence bearing the V2 and J1–3 regions. The hypervariable sequence between V and J forms a bulge structure after hybridization with the M-Probe, with the bulge size of 26 nt. **d**, Summary of observed fluorescence for the M-Probes after an overnight hybridization reaction. The main diagonal corresponds to on-target hybridization in which the V and J regions of the target are matched to the M-Probe segments, and off-diagonal blue squares correspond to hybridization reactions in which one of the V and J regions differs between the M-Probe and the target. Hybridization reactions in which the target and M-Probe differ in both the V and J regions were not tested (grey squares). All experiments were performed in triplicate at 37 °C in 1× PBS, [M-Probe] = 100 nM, [Target] = 300 nM. **e**, Distribution of observed off-target fluorescence (left) and on-target fluorescence (right).

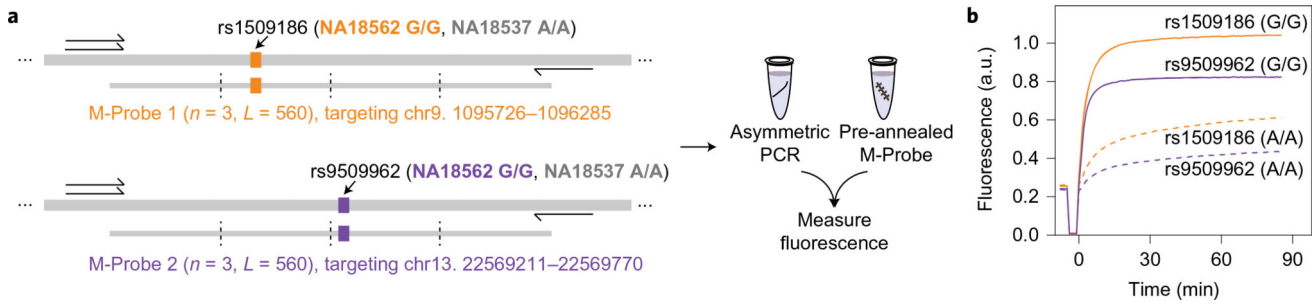

Figure 4. M-Probe detection of single nucleotide variants within 560 nt targeting regions

a, The rs1509186 SNP is homozygous G/G in the NA18562 genomic DNA (gDNA) sample and homozygous A/A in NA18537. PCR primers were designed to amplify a 590 nt amplicon containing the rs1509186 SNP and an $n = 3$ M-Probe was designed to bind to 560 nt of the amplicon with sequence-matching NA18562. No other sequence differences between NA18537 and NA18562 are expected in the 560 nt targeting region, based on the 1000 Genomes Project database. The SNP is located 196 nt from the 5′ end of the M-Probe target region. Similarly, a separate 590 nt amplicon is generated around the rs9509962 SNP; this second M-Probe is also 560 nt long and the SNP lies 285 nt from the 5′ end of the target region. Vertical dashed lines denote junctions separating M-Probe segments. Supplementary Section 5 shows probe and target preparation details. **b**, Fluorescence responses of the M-Probes (10 nM final concentration) to respective amplicon targets. Hybridization experiments were performed at 45 °C in 1× PBS. Amplicons from the NA18562 gDNA sample (solid lines) induced significantly higher fluorescence than amplicons from the NA18537 sample (dashed lines), indicating that the M-Probes are selective for even single nucleotide variants across a 560 nt target sequence.

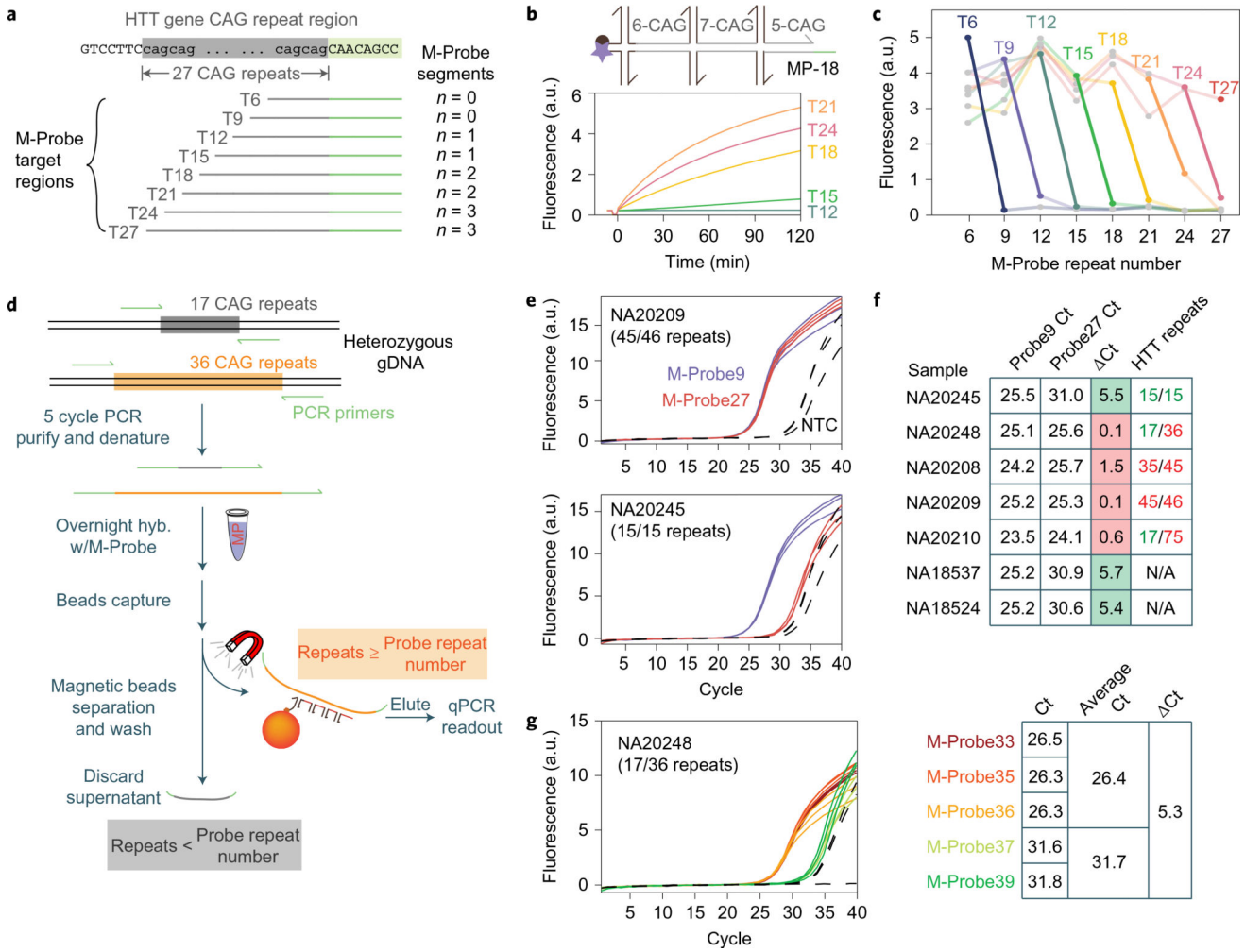

Figure 5. Profiling triplet repeat number using M-Probes

a, A series of M-Probes were designed, each targeting a threshold number of triplet repeats (CAG in the case of Huntington's gene, HTT). Only target sequences meeting or exceeding its threshold repeat number will hybridize significantly to an M-Probe. In addition to the repeat region (yellow), the M-Probe also binds to an 8 nt downstream sequence (green) to ensure specific hybridization to the HTT gene. **b**, Experimental fluorescence response of synthetic oligonucleotide targets bearing different numbers of triplet repeats to an M-Probe designed to detect targets with ≥18 CAG repeats. Here, [M-Probe] = 10 nM and [Target] = 30 nM; hybridization proceeded at 37 °C in 1× PBS. **c**, Summary of M-Probe responses to eight synthetic targets. Minimal hybridization was observed when the target repeat number was less than the M-Probe repeat number. **d**, Workflow for selective capture of high repeat HTT gene from genomic DNA using biotin-functionalized M-Probes (Supplementary Fig. 6–8). Streptavidin-functionalized magnetic beads were used to separate bound from unbound DNA. Captured DNA molecules were subsequently amplified and quantitated by qPCR. **e**, qPCR amplification traces of captured HTT gene from the NA20209 and NA20245 genomic DNA samples (125 ng gDNA initial input per reaction). M-Probes targeting 9 repeats (M-Probe9) and 27 repeats (M-Probe27) were used to classify gDNA samples. The

five cycle threshold difference (ΔCt) observed for NA20245 indicates that capture of HTT genes with above the threshold number of repeats was roughly 30-fold more efficient than for those below the threshold. **f**, Summary of experimentally observed Ct values for seven genomic DNA samples (mean values for triplicate runs). Samples with fewer than 27 repeats show more than five-cycle ΔCt, and samples with expanded triplet repeats (at risk for Huntington's disease) exhibit less than two-cycle ΔCt. **g**, Precise determination of triplet repeat number using a series of M-Probes. Five different M-Probes targeting 33, 35, 36, 37 and 39 repeats were constructed and applied to the NA20248 genomic DNA sample. NA20248 was determined correctly to possess an HTT gene with 36 repeats based on the observed Ct values.