|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 1:** Cas12a crRNAs and bottom strand (BS) barcodes in this study. | | | | | | | | | | | |
| **BS barcode 5'->3'** | /5Phos/AGGTGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/CCTGGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/CTGCGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/CTGCGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/GAGAGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/GTGTGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/GTCTGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/TCACGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/CAAGGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/TCTGGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT | /5Phos/CTCCGGTGCTGAAGAAAGTTGTCGGTGTCTTTGTGTTAACCTTAGCAAT |
| **overhang** | AGGT\_- | CCTG\_- | CTGC\_+ | GAGG\_+ | GAGA\_+ | GTGT\_+ | GTCT\_+ | TCAC\_- | CAAG\_- | TCTG\_+ | CTCC\_+ |
| **protospacer 5'->3'** | CAGCCTTCCTTCCACACGCAC | TAACTTTATCTTTCCTTAACA | ATGGAAACCAAGGGCCAAGGC | AGCCCTATTGGGTTCTTGGCC | ACTTCCGGTGGAGGGCCGCCT | CAGTACCAGAAAGTTCACAAC | TCACAGTTCCAAGTTTCTCAG | TTCCTCCCTTTCTTCCTCGGT | CCACCCTCTCTCCCCACTACT | CTAAAGTGGCAGGCCTTGGCA | AGCAAGTCTGTGTCATCTCGG |
| **strand** | - | - | + | + | + | + | + | - | - | + | + |
| **hg38 ref** | chrX: 147913437-147913457 | chrX: 147916120-147916140 | chrX: 147908319-147908339 | chrX: 147910467-147910487 | chrX: 147911951-147911971 | chr9: 27571877-27571897 | chr9: 27571964-27571984 | chr9: 27574580-27574600 | chr9: 27573669-27573689 | chr9: 27572283-27572303 | chr9: 27572809-27572829 |
| **crRNA name** | FMR1\_Cpf1\_147913437\_C1 | FMR1\_Cpf1\_147916120\_C2 | FMR1\_Cpf1\_147908319\_C3 | FMR1\_Cpf1\_147910467\_C4 | FMR1\_Cpf1\_147911951\_C5 | C9orf72\_Cpf1\_27571877\_C1 | C9orf72\_Cpf1\_27571964\_C2 | C9orf72\_Cpf1\_27574580\_C3 | C9orf72\_Cpf1\_27573669\_C4 | C9orf72\_Cpf1\_27572283\_C5 | C9orf72\_Cpf1\_27572809\_C6 |
| **crRNA ID** | FMR1-C1 | FMR1-C2 | FMR1-C3 | FMR1-C4 | FMR1-C5 | C9orf72-C1 | C9orf72-C2 | C9orf72-C3 | C9orf72-C4 | C9orf72-C5 | C9orf72-C6 |
| **ROI** | FMR1 | FMR1 | FMR1 | FMR1 | FMR1 | C9orf72 | C9orf72 | C9orf72 | C9orf72 | C9orf72 | C9orf72 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 2:** Cas9 crRNAs used in this study. | | | | | | | | | |
| **protospacer 5'->3'** | CCACTTGAAGAGAGAGGGCG | ACAGCGTTGATCACGTGACG | GATTAAGGCAGCTATAAGCA | GTTGAGGAAAGGCGAGTACG | CATCCTGATCCTAATAAAAG | CCCCGGGAAGGAGACAGCTC | AGATAGACCCAATGAGCACA | AAACTGGTCTCAGGTCACAA | TCCATAAGCTGTGAAGCCGG |
| **strand** | + | + | + | + | - | + | + | - | - |
| **hg38 ref** | chrX: 147911806-147911825 | chrX: 147911858-147911877 | chrX: 147910985-147911004 | chrX: 147911229-147911248 | chrX: 147932675-147932694 | chr9: 27573134-27573153 | chr9: 27572706-27572725 | chr9: 27574815-27574834 | chr9: 27576480-27576499 |
| **crRNA name** | FMR1\_Cas9\_147911806\_C1.1 | FMR1\_Cas9\_147911858\_C2.1 | FMR1\_Cas9\_147910985\_C3.1 | FMR1\_Cas9\_147911229\_C4.1 | FMR1\_Cas9\_147932675\_C5.1 | C9orf72\_Cas9\_27573134\_C1.1 | C9orf72\_Cas9\_27572706\_C2.1 | C9orf72\_Cas9\_27574815\_C.3.1 | C9orf72\_Cas9\_27576480\_C4.1 |
| **crRNA ID** | FMR1-C1.1 | FMR1-C2.1 | FMR1-C3.1 | FMR1-C4.1 | FMR1-C5.1 | C9orf72-C1.1 | C9orf72-C2.1 | C9orf72-C3.1 | C9orf72-C4.1 |
| **ROI** | FMR1 | FMR1 | FMR1 | FMR1 | FMR1 | C9orf72 | C9orf72 | C9orf72 | C9orf72 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Supplementary Table 3:** Flow cell throughput and enrichment statistics for whole genome sequencing. | | | |
| run | FAK68900 | PAD42311 | PAD43667 |
| sample | 24/5#2 | 24/5#2 | 24/5#2 |
| nt\_c9orf72 | 26461 | 25834 | 27006 |
| nt\_fmr1 | 138118 | 139129 | 159991 |
| nt\_genome | 26231250257 | 20050285547 | 23313676905 |
| d | 15,628 | 15,665 | 17,757 |
| s | 6,27E-06 | 8,23E-06 | 8,02E-06 |
| ef | 1,715 | 2,249 | 2,193 |
| **nt\_c9orf72**: Sequenced Bp on c9orf72 target. **nt\_fmr1**: Sequenced Bp on FMR1 target. **nt\_genome**: Sequenced Bp excluding enrichment targets. **d**: Mean coverage on enrichment targets (nt / length). **s**: Specificity as (Bp on target) / (Bp off target). **ef**: Enrichment as (mean coverage on target) / (mean coverage genome).  ROI sizes: C9orf72 – 2599 Bp; FMR1 – 5102 Bp | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 4:** Flow cell throughput and enrichment statistics for overhang protocol | | | | | | | | | | | |
| e | NA | NA | NA | 0,29 | 0,18 | 0,40 | 0,00 | 0,00 | Inf | 0,60 | 0,17 |
| ef | 98,01 | 171,51 | 477,58 | 2760,78 | 1604,67 | 1410,75 | 169,78 | 344,00 | 244,25 | 43,55 | 187,12 |
| s | 2,69E-04 | 4,71E-04 | 1,31E-03 | 2,52E-03 | 1,46E-03 | 1,29E-03 | 1,55E-04 | 3,14E-04 | 2,23E-04 | 3,97E-05 | 1,71E-04 |
| d | 1,14 | 5,75 | 0,95 | 5,12 | 11,86 | 6,04 | 0,73 | 0,60 | 1,62 | 3,74 | 5,85 |
| exp | 0 | 10 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 3 | 1 |
| wt | 0 | 0 | 0 | 7 | 11 | 5 | 1 | 1 | 0 | 5 | 6 |
| n\_valid | 0 | 10 | 2 | 9 | 13 | 7 | 1 | 1 | 2 | 8 | 7 |
| n | 1 | 13 | 3 | 11 | 20 | 9 | 2 | 1 | 3 | 8 | 11 |
| nt\_genome | 33505853 | 96491100 | 5711508 | 5337747 | 21270094 | 12320787 | 12350426 | 5037174 | 19069191 | 246888420 | 90016304 |
| nt\_fmr1 | 9018 | 45448 | 7491 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nt\_c9orf72 | 0 | 0 | 0 | 13437 | 31122 | 15849 | 1912 | 1580 | 4247 | 9805 | 15359 |
| crRNA | C3,C5 | C3 | C1,C3 | C2,C4 | C4 | C3,C4 | C1,C4 | C3,C4 | C1,C4 | C4 | C4 |
| Sample | iPS7 | iPS7 | iPS6 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 |
| run | FAH51749 | FAH65171 | FAH66274 | FAH67301 | FAH67421 | FAH72446 | FAH80338 | FAH81424 | FAH81473 | FAH87258 | FAH88643 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 5:** Flow cell throughput and enrichment statistics for Cas12 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| e | 0,86 | Inf | 0,71 | NA | Inf | 2,50 | Inf | 0,00 | 1,00 | 5,00 | 0,75 | 3,36 | 0,69 | 4,50 | 0,50 | 4,18 | 0,77 | 2,80 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| ef | 28,92 | 28,92 | 30,33 | 30,33 | 381,72 | 7,31 | 7,31 | 0,98 | 11,84 | 11,84 | 47,39 | 47,39 | 73,74 | 73,74 | 235,36 | 235,36 | 25,89 | 25,89 | 40,01 | 40,01 | 79,00 | 79,00 | 152,60 | 152,60 | 195,69 | 195,69 | 15,73 | 15,73 | 237,26 | 237,26 |
| s | 1,06E-04 | 1,06E-04 | 1,11E-04 | 1,11E-04 | 3,48E-04 | 2,67E-05 | 2,67E-05 | 3,59E-06 | 4,33E-05 | 4,33E-05 | 1,73E-04 | 1,73E-04 | 2,70E-04 | 2,70E-04 | 8,61E-04 | 8,61E-04 | 9,47E-05 | 9,47E-05 | 1,46E-04 | 1,46E-04 | 2,89E-04 | 2,89E-04 | 5,58E-04 | 5,58E-04 | 7,16E-04 | 7,16E-04 | 5,76E-05 | 5,76E-05 | 8,68E-04 | 8,68E-04 |
| d | 17,78 | 17,78 | 21,13 | 21,13 | 9,08 | 4,83 | 4,83 | 0,34 | 20,47 | 20,47 | 61,62 | 61,62 | 44,38 | 44,38 | 296,80 | 296,80 | 19,96 | 19,96 | 7,10 | 7,10 | 59,41 | 59,41 | 34,90 | 34,90 | 247,94 | 247,94 | 11,37 | 11,37 | 143,91 | 143,91 |
| exp | 31 | 0 | 36 | 0 | 1 | 5 | 0 | 0 | 9 | 0 | 40 | 0 | 50 | 0 | 146 | 0 | 53 | 0 | 0 | 0 | 0 | 2 | 0 | 4 | 1 | 166 | 0 | 11 | 0 | 147 |
| wt | 36 | 3 | 51 | 0 | 0 | 2 | 5 | 1 | 9 | 22 | 53 | 55 | 72 | 10 | 290 | 491 | 69 | 90 | 21 | 0 | 291 | 0 | 65 | 0 | 564 | 0 | 0 | 0 | 261 | 0 |
| n\_valid | 67 | 3 | 87 | 0 | 1 | 7 | 5 | 1 | 18 | 22 | 93 | 55 | 122 | 10 | 436 | 491 | 122 | 90 | 21 | 0 | 291 | 2 | 96 | 4 | 565 | 166 | 0 | 11 | 261 | 147 |
| n | 74 | 5 | 139 | 2 | 14 | 10 | 5 | 1 | 22 | 25 | 102 | 63 | 228 | 19 | 601 | 623 | 152 | 102 | 28 | 1 | 350 | 4 | 151 | 5 | 779 | 317 | 2 | 28 | 283 | 181 |
| target | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | c9orf72 | fmr1 | c9orf72 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 |
| nt\_genome | 1770275560 | 1770275560 | 2005283552 | 2005283552 | 68441046 | 1903520306 | 1903520306 | 996208602 | 4977573462 | 4977573462 | 3743310604 | 3743310604 | 1732575557 | 1732575557 | 3630396753 | 3630396753 | 2219834694 | 2219834694 | 510929403 | 510929403 | 2164853244 | 2164853244 | 658391445 | 658391445 | 3647449835 | 3647449835 | 2079599550 | 2079599550 | 1746227088 | 1746227088 |
| nt\_fmr1 | 41756 | 41756 | 2426 | 2426 | 0 | 30101 | 30101 | 754 | 170407 | 170407 | 432826 | 432826 | 110184 | 110184 | 2023404 | 2023404 | 150005 | 150005 | 7264 | 7264 | 25602 | 25602 | 24091 | 24091 | 782044 | 782044 | 116105 | 116105 | 812615 | 812615 |
| nt\_c9orf72 | 145516 | 145516 | 220085 | 220085 | 23822 | 20774 | 20774 | 2819 | 45190 | 45190 | 216089 | 216089 | 357202 | 357202 | 1102213 | 1102213 | 60201 | 60201 | 67523 | 67523 | 600010 | 600010 | 343438 | 343438 | 1828983 | 1828983 | 3581 | 3581 | 702935 | 702935 |
| crRNA | C1-C4 | none | C1-C4 | none | C1-C4 | C1-C3 | C1-C4 | C1-C3 | C1-C3 | C1-C4 | C1-C3 | C1-C4 | C1-C3 | C3 | C1-C3 | C1-C4 | C1-C3 | C1-C4 | C1-C4 | C1-C4 | C1-C4 | C3 | C1-C3 | C3 | C1-C3 | C1-C4 | none | C1,C3 | C1-C3 | C1-C4 |
| sample | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | iPS6 | iPS6 | iPS6 | iPS6 | iPS6 | iPS6 | iPS6 | iPS6 | iPS7 | iPS7 | iPS7 | iPS7 |
| run | FAH91937 | FAH91937 | FAJ02524 | FAJ02524 | FAJ03272 | FAK02383 | FAK02383 | FAK02402 | FAK57423 | FAK57423 | FAK67802 | FAK67802 | PAD01039 | PAD01039 | PAD01249 | PAD01249 | PAD42366 | PAD42366 | FAH66294 | FAH66294 | FAJ02378 | FAJ02378 | PAD01034 | PAD01034 | PAD01413 | PAD01413 | FAK02017 | FAK02017 | FAK58936 | FAK58936 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 6:** Flow cell throughput and enrichment statistics for Cas9 protocol | | | | | | | | | | | | | | | | | | |
| e | 0,630 | 2,307 | 0,763 | 3,778 | 0,000 | Inf | 0,824 | 6,143 | 0,622 | 4,250 | 0,707 | 3,069 | 0,537 | 6,545 | 0,537 | 4,209 | 0,594 | 2,814 |
| ef | 900,89 | 900,89 | 1640,69 | 1640,69 | 305,52 | 305,52 | 323,16 | 323,16 | 1211,78 | 1211,78 | 1328,34 | 1328,34 | 1070,81 | 1070,81 | 1278,74 | 1278,74 | 714,28 | 714,28 |
| s | 3,30E-03 | 3,30E-03 | 6,00E-03 | 6,00E-03 | 1,12E-03 | 1,12E-03 | 1,18E-03 | 1,18E-03 | 4,43E-03 | 4,43E-03 | 4,86E-03 | 4,86E-03 | 3,92E-03 | 3,92E-03 | 4,68E-03 | 4,68E-03 | 2,61E-03 | 2,61E-03 |
| d | 204,166 | 204,166 | 29,656 | 29,656 | 0,913 | 0,913 | 28,809 | 28,809 | 42,419 | 42,419 | 73,089 | 73,089 | 66,514 | 66,514 | 328,341 | 328,341 | 136,491 | 136,491 |
| exp | 179 | 0 | 29 | 0 | 0 | 0 | 14 | 0 | 23 | 0 | 53 | 0 | 44 | 0 | 309 | 0 | 111 | 0 |
| wt | 284 | 233 | 38 | 39 | 2 | 1 | 17 | 47 | 37 | 59 | 75 | 109 | 82 | 79 | 575 | 319 | 187 | 154 |
| valid | 463 | 233 | 67 | 39 | 2 | 1 | 31 | 47 | 60 | 59 | 128 | 109 | 126 | 79 | 884 | 319 | 298 | 154 |
| n | 639 | 316 | 94 | 46 | 2 | 1 | 36 | 56 | 73 | 69 | 157 | 127 | 151 | 100 | 1041 | 367 | 364 | 197 |
| target | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 | c9orf72 | fmr1 |
| nt\_genome | 652420919 | 652420919 | 52036368 | 52036368 | 8599520 | 8599520 | 256637751 | 256637751 | 100774507 | 100774507 | 158401558 | 158401558 | 178821500 | 178821500 | 739197596 | 739197596 | 550117330 | 550117330 |
| nt\_fmr1 | 1358873 | 1358873 | 201194 | 201194 | 4956 | 4956 | 257529 | 257529 | 343463 | 343463 | 561077 | 561077 | 475210 | 475210 | 1977759 | 1977759 | 941359 | 941359 |
| nt\_c9orf72 | 791198 | 791198 | 111117 | 111117 | 4655 | 4655 | 45854 | 45854 | 103248 | 103248 | 208621 | 208621 | 225250 | 225250 | 1480001 | 1480001 | 496032 | 496032 |
| crRNA | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 | C1.1-C4.1 | C1.1-C5.1 |
| sample | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 | 24/5#2 |
| run | FAJ04502 | FAJ04502 | FAJ04588 | FAJ04588 | FAK01734 | FAK01734 | FAK01877 | FAK01877 | FAK57718 | FAK57718 | FAK58137 | FAK58137 | FAK62030 | FAK62030 | FAK67994 | FAK67994 | PAD43259 | PAD43259 |

|  |  |
| --- | --- |
| **Supplementary Table 7:** Oligomere sequences used for synthetic (G4C2)n-repeat cloning. | |
| Oligo ID | Oligo sequence 5’→3’ |
| 3x GGGGCC fwd | GGATCCGCTCTTCCGGGGCCGGGGCCGGGGCCGGGGCCTGCGGCCGC |
| 3x GGGGCC rev | GCGGCCGCAGGCCCCGGCCCCGGCCCCGGCCCCGGAAGAGCGGATCC |
| 4x GGGGCC fwd | GGATCCGCTCTTCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCTGCGGCCGC |
| 4x GGGGCC rev | GCGGCCGCAGGCCCCGGCCCCGGCCCCGGCCCCGGCCCCGGAAGAGCGGATCC |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Supplementary Table 8:** Comparison of methods for the determination of repeat expansion status and quantification of repeat expansion counts (\* this study; rn, repeat number) | | | | |
| *Technology* | *targeting strategy* | *resolution* | *single molecule* | *targeting/ coverage* |
| *Southern blot* | hybridization probe | + low rn | no | +++ |
| (+) high rn |
| *repeat primed PCR* | PCR primer | +++ low rn | no | +++ |
| - high rn |
| *Sanger* | Sequencing primer | +++ low rn | no | +++ |
| - high rn |
| *Illumina short read* | n.a. | +++ low rn | no | +++ |
| - high rn |
| *PacBioNo-Amp SMRT-seq* | Cas9-targeting of region, hairpin adaptor ligation and biotin pulldown | +++ low rn | yes | (+) |
| (+) high rn |
| *Nanopore*  *Cas12a-RNP\** | Cas12a-RNP targeting, and selective adaptor ligation | +++ low rn | yes | +++ |
| +++ high rn |

## Supplementary Note

### 1 Considerations pertaining to the nomenclature, selection of prototypical examples for unstable repeat expansions.

In our study, we set out to explore this technology as means to specifically sequence selected genomic regions, to ascertain repeat expansion status and to determine CpG methylation patterns on the single DNA molecule level from cells from c9FTD/ALS and FXS patients as prototypical examples for STR expansion disorders.

We acknowledge that several, at times contradictory, abbreviations and names have been used for unstable repeat expansions in health and disease.

The term “Short Tandem Repeat” (STR) have been recently become more prevalently used in the context of disorders of unstable repeat expansions (see for some recent examples 1-3).

While writing the manuscript, our co-authors with background in clinical human genetics pointed out that the term ‘STR’ has been also previously used in the context of so called microsatellites, numerous polymorphic genomic loci which are not necessarily associated with any disorder or pathology 4.

Both methods described in our study are generally applicable to any genomic region (CRISPR Cas12a-/Cas9-RNP based enrichment) or any STR (STRique) independently of their potential pathogenic role. Therefore, we took a comprehensive stance for the writing of this manuscript and use the terms STR and repeat expansions interchangeably. We note that depending on resource consulted or nomenclature followed, the term “tandem repeat” and associated concepts (short tandem repeat, variable nucleotide tandem repeat, microsatellite, minisatellite) are frequently used with overlapping and at times inconsistent ways in the existing literature.

For consistency, we semantically follow the structure of the Medical Subject Header (MeSH) descriptor data 2018 structure for Tandem Repeat Sequence [G05.360.340.024.850, <https://meshb.nlm.nih.gov/record/ui?ui=D020080>] and use the terms ‘Short Tandem Repeats’ (STRs) interchangeably with “Microsatellite Repeats” characterized by a short repeat unit of 2-8 base pairs that is tandemly repeated (see Suppl. Fig. 11).

We also note, that in contrast most several recent studies refer to STRs as consisting of 1-6 bp motifs (see for example Hannan, Nat Rev Genetics 2018 5 and Dashnow et al. Genome Biology 2018 1). The frequently used term “Variable Nucleotide Tandem Repeats” (VNTR) is suggested by MeSH as alternative term for minisatellite repeats [G05.360.340.024.850.550]. We acknowledge that not all MeSH descriptors may reflect current consensus in the field or expert knowledge and may lack comprehensiveness (e.g. tandem repeats with nine base pairs as repetitive motiv would be - following the MeSH descriptors - neither microsatellites nor minisatellites nor STRs nor VNTRs ), but we chose the MeSH ontology as actively curated, maintained and publicly available resource for a systematic view on repetitive DNA sequences.

To give an overview of the conceptual space of tandemly repeated DNA sequence motives, Suppl. Fig. 11 schematically depicts the MeSH descriptor terms with visual representations of n-tuple sequence motifs.

For context, we have added all four currently available studies (as of 2018/10/15) that have used nanopore sequencing to characterize different tandem repeats on the background of the human genome and added the terms used in these studies in the context of MeSH descriptors. On the right side of Supplementary Fig. 11 we list technical challenges specific to the application of nanopore sequencing to the different subsets of tandem repeats.

Importantly, different repeat expansion classes pose different technical problems from the perspective of using third generation sequencing technologies (Suppl. Table 8). In our study we experimentally focused on STRs with 3-6 bp repetitive elements as these pose a nanopore sequencing and signal processing challenge distinct from those associated with VNTRs/minisatellites or even larger tandem repeat structures such as centromeric repeats and will accommodate for most currently known and clinically relevant tandem repeat expansion associated disorders.

Briefly, nanopore signals from 3-6 bp STRs deviate strongly from many assumptions and parameter sets used for general purpose base callers (such as ONT’s Albacore and Guppy). For example, as can be seen from Fig. 1A and Suppl. Fig. 2, raw signals from repeats are much more homogenous and often are of smaller amplitude when compared to more variable sequences. In comparison, the much larger VNTRs/minisatellites can be effectively basecalled with general purpose base callers such as Albacore but base-calling error rates in single reads of 10-15 % pose a set of problem for the accurate determination of repeat expansion counts 6 different from those encountered with shorter repetitive motifs. Very large tandem repeat units, in contrast, can be more effectively base-called and sequence variants can be detected through raw signal polishing methods but reads spanning several hundreds of kilobases are necessary to correctly assemble such higher order repetitive sequences.

From a clinical perspective, unstable genomic short repetitive sequences (STRs with 3-6 bp) cause more than 30 Mendelian human disorders 5,7,8. In this study, we employ FXS and c9FTD/ALS as prototypical examples for clinically relevant repeat expansion disorders in humans to demonstrate the general applicability and potential usefulness of the combination of both novel methods towards an integrative analysis of unstable repeat expansions for basic research and clinical applications.

We chose the GGGGCC hexanucleotide [(G4C2)n] repeat expansion in the C9orf72 gene as an example for a hexanucleotide repeat expansion disorders, as it represents the first and most common known genetic variant underlying both Amyotrophic Lateral Sclerosis and Frontotemporal Dementia (c9FTD/ALS) 9,10. Based on Southern blot analyses, (G4C2)n-expansions in c9FTD/ALS patients have been estimated in the range of 100 - 7500 repeats resulting in 0.6 to 45 kB regions with 100% Guanine/Cytosine content and G-quadruplex formation 11-13. These sequence features and structural DNA variants are often mentioned as reason, why conventional, polymerase-based molecular methods fail to resolve repeat expansion numbers higher than 100 - 150 repeats 9,10,13,14.

We also choose Fragile X Syndrome (FXS) as second example for a STR expansion disorder. In FXS, as it represents the first described repeat expansion disorder, the accumulation of more than 200 CGG-repeats impair the function of the FMR1-gene and subsequently result in neurodevelopmental delay, intellectual disability and frequently autism 15.

In both expansion disorders, STR expansions have been shown to be intra- and inter-individually highly variable with suggestive evidence supporting ageing, genetic anticipation, neural differentiation and epigenetic mechanisms as underlying this phenomenon: Strong evidence exists for genetic anticipation and in c9FTD/ALS and FXS 5,16. Genetic anticipation describes the fact that disease severity in most if not all STR expansion disorders increases through successive generations of affected family members, i.e. progressively leads to earlier onset and more severe phenotypes. Genetic anticipation is now mainly ascribed to the observation that expanded repeats dynamically increase their length over generations and that longer repeats tend to be increasingly unstable and appear to be associated with more severe phenotypes. Estimates of c9FTD/ALS-STR length have been associated with patient’s age, age of disease onset, disease severity and vary between tissues and brain regions 12,17-22.

In both c9FTD/ALS and FXS, increased methylation of CpG islands close to the expanded STRs have been also implicated to affect disease phenotype 19,20,23-25.

We note, that due to the focus of this study on the universal nature of our proposed approach and inherent space constraints, we can only superficially describe the clinical presentation and specific aspects of both disorders, such as premutation status, associated disorders such as Fragile X associated tremor/ataxia syndrome (FXTAS) in the case of female FMR1-CGG expansion carriers.

In this discussion, we therefore highlight only aspects of both disorders pertinent to the interpretation and novel insights enabled by the methods. We encourage interested readers to consult existing, excellent reviews 5,8,16 as a starting point for a much more comprehensive summary of unstable repeat expansion disorders in general and c9FTD/ALS and FXS in particular.

### 2 Considerations pertaining to the current challenges for exact repeat expansion quantification with polymerase based methods, Southern blotting and third generation sequencing methods

As of now, most methods for the determination of repeat expansion status each suffer from at least two of the following three problems: i) lack of resolution at higher repeat numbers, i) lack of single molecule analysis capability and/or iii) targeting of expansion loci for sufficient coverage of higher repeat expansion counts.

Because of the limitations of each currently available methods, the combination of PCR and Southern blots is still recommended in the diagnostic setting (e.g. for FXS). 26,27

Conventional, polymerase-based methods (PCR-reactions, Sanger-sequencing) are currently frequently used to estimate repeat numbers, but are imperfect beyond the threshold of 100-150 repeats 1,10,28 as DNA polymerases, e.g., in conventional PCR or sequencing reactions cannot read through the repetitive, GC-rich and often secondary structures forming sequences. Although, while few heterogeneous alleles with lower repeat counts can be detected with PCR-methods, this readout degrades quickly with higher repeat counts > 100. Also, no conventional polymerase-based methods for repeat expansion status on the single molecule level is currently available.

Short read sequencing technologies (mainly Illumina) limited by their read/insert length29 as even paired end reads can span only smaller repeat expansions and synthetic long read methods coupled with Illumina short read sequencing also cannot resolve longer repeat expansions as the mapping problem with sequences containing 100 % repetitive elements remain.

Southern blot analysis requires large amounts of DNA and cannot be scaled down to single molecule level. Additionally, the resolution of gel electrophoresis analysis decreases strongly with repeat expansion length (see for example Suppl. Fig. 9a - h and here specifically the standard curves Suppl. Fig. 9d for the c9FTD/ALS and Suppl. Fig. 9h). We note, that only Southern blots are currently realistically capable of resolving intraindividual repeat instability/heterogeneity if different distinct high repeat number expansion maxima exist as has been discussed e.g. in Almeida et al. 201330. We also highlight the fact that currently established diagnostic Southern Blot workflow require between 3 - 4 days of processing time, while a nanopore sequencing based workflow could enable next day or even same day diagnostic results (see Suppl. Fig. 12).

Current third generation single molecule sequencing methods (PacBio SMART-seq and ONT nanopore sequencing) have been consistently shown to be capable to read through long stretches of repetitive DNA from tandem repeat expansion disorder derived plasmids31-34, but their throughput from patient samples has been up to now severely limited.

As illustrative example, Ebbert and colleagues32 recently reported the results from a whole genome SMRT-seq experiment employing five PacBio flow cells with a total of four PacBio reads covering C9orf72 repeat expansion region (30-1324 repeats). Only one read with 1324 repeat counts fully bridged the repeat expansion region. A targeted enrichment experiment utilizing the “No-Amp” Cas9-targeting SMRT-seq technology 35 in the same study resulted in a total of 134 reads from expanded alleles with a repeat expansion count distribution (two modes at 110 and 870), which were interestingly not matching results obtained from Southern blot analysis from the same subject (around estimated 1000 repeats in the Southern blot). Ebbert et al. also reported the results from nanopore whole genome sequencing experiments with 2 reads from the wild type allele and no single read covering the C9orf72 repeat expansion could be recovered from 15 flow cells loaded with whole genome libraries.

### 3 Considerations pertaining to reporting enrichment results

In this study we report our enrichment results in reads per target per individual flow cell which equals a single DNA sample from a single individual. From this metric, we derive sub-counts, e.g. reads covering the target with or without repeat expansion and from those, which appear to contain a repeat expansion, those, from which we could derive a repeat count with STRique.

We believe that this metric captures best the real-world results, where a diagnostician or researcher analyzes a sample from an individual to learn about the length of a possible repeat expansion and its epigenetic context in one flow cell. For consistency reasons we chose to report only these results in the main text.

We note that a set of several other metrics have been proposed previously to describe targeted sequencing enrichment results 36. The parameters proposed were adapted to our specific experimental setup and research question as described below the parameter description written in bold.

**(i) Region of interest (size): ROIwt;**

here we report the fragment size of the wild type allele between the two most proximal crRNA binding sites

* **(ii) Total average read depth (in ROI): D;**

here we report the mean coverage over the targeted repeat expansion site

* **(iii) Fraction of ROI sufficiently covered (at a specified D): F;**

NA

* **(iv) Specificity (coverage of ROI): S;**

here we report the sequenced Bp across all targeted ROIs versus the sequenced Bp of the background genome excluding the ROI

* **(v) Enrichment Factor (D for ROI versus D): EF;**

here we report the mean coverage across all targeted ROIs versus the mean coverage of the genome excluding the ROIs

* **(vi) Evenness (lack of bias): E;**

here we report in case of two alleles with one wild type and the other  
 expanded the ratio expanded allele over the wild type allele

* **(vii) Weight (input DNA requirement): W**

here we report the actual amount of DNA used for a specific experiment

Results from all flow cells following the above metrics are listed in Supplementary Table 3-6. Apart from these theoretical considerations, we are taking a pragmatic approach to the question of how to conceptualize the molecular method described in our study:

With whole-genome sequencing on the current generation MinION fow cells, we can obtain an average of about 10 reads from any genomic target region. Using programmable nucleases, we can increase the number of reads obtained up to a hundred times from genomic regions with a size 5 - 20 kb (i.e., up to 1000 reads; see Main Fig. 2). The higher number of reads on target enables us to answer clinically and scientifically relevant questions about these regions by increasing the information content obtained in a single experiment (flow cell) selectively for these regions, while we significantly decrease the amount of information gathered from the rest of the genome. In light of these considerations, we believe it may be acceptable to speak of "enrichment" in the sense of "enrichment of information" obtained from a distinct genomic region in a nanopore experiment.

### 4 Considerations pertaining to the value of a Hidden Markov Model -based repeat quantification algorithm instead of a neural networks base calling algorithm for the quantification of unstable repeat expansions in nanopore reads

For this study we have developed an algorithm for a more exact determination of repeat expansion numbers from nanopore raw traces because, in our evaluation of results obtained from general purpose base-calling implementations, accurate STR number estimation beyond approximately ten (G4C2)n-repeats became highly unreliable (Suppl. Fig. 4). We also attempted to use previously published repeat expansion quantification algorithms (RepeatHMM, STRetch), which can use noisy, long-read base space level data to estimate repeat expansion numbers (Fig. 1B and Suppl. Fig. 3). For this we initially used the generally best performing basecaller at the time, Albacore v2.3.3 developed by Oxford Nanopore Technologies (ONT). In the meantime, significantly better performing general purpose base-calling algorithms have been made publicly available by ONT (Guppy-flipflop). While the manuscript was under review, Albacore was deprecated by ONT.

The significant improvements in base calling performance of Guppy over Albacore are mainly due to the use of more sophisticated deep learning methods for model training. Wick, Judd, and Holt recently compared the performance of all currently available neural network basecallers for Oxford Nanopore sequencing 37 and provided a detailed discussion of the underlying technology. In the context of our manuscript, we highlight the point that a general purpose base caller's performance build with deep-learning strategies on native DNA strongly depends on the data set used to train its model as well as the complexity of the neural network architecture employed 37.

We also note that the representation of sequence features in a training dataset is particularly crucial for the performance of a resulting basecalling model. Specifically, a basecaller performs better on native DNA (i.e., DNA containing base modifications such as cytosine methylation) if a model is applied to samples stemming from organisms of the same taxon as the model has been trained on37. This increase in performance has been attributed to the incorporation of taxon-specific DNA methylation patterns into the basecaller model.

Important for training basecaller, (G4C2)n-expansions are not a common feature in the human genome. A recent study estimated that about 344 loci with two consecutive (G4C2)n-sequences are present in the human reference genome, and less than 10 of those loci in proximity to genomic candidate regions for Amyotrophic Lateral Sclerosis identified in GWAS studies displayed variable repeat lengths up to 15 repeats in a cohort of ALS patients 38. Based on our and others39 difficulties with stably propagating even moderately sized (G4C2)n-repeat expansions in bacterial cells, we posit that k-mers containing repetitive (G4C2)n-sequences may also be likely substantially underrepresented in most if not all datasets from all eukaryotic and prokaryotic organisms that have been used so far for training nanopore base calling models.

This assumption is supported by our finding that in the particularly challenging case of (G4C2)n-repeat expansions, the current versions of Guppy-flipflop (May 2019) in combination with either a decoy alignment strategy (similar to the STRetch algorithm 1) or RepeatHMM 40 for repeat quantification are outperformed by STRique.

Interestingly, Guppy-flipflop in high-accuracy mode with a window size of 16 kb in combination RepeatHMM works reasonably well on a subset of expanded (G4C2)-sequences from the - strand of the locus (see Supplementary Methods). This finding indicates that future versions of neural network base callers may have the potential to perform equally well for this particularly repeat expansion.

In other clinically relevant STR expansion disorders, such as Huntington's disease (HD; OMIM: # 143100) the expansion length is limited to rare cases with a maximum of 120 consecutive CAG-repeats, and the average repeat expansion length in affected individuals is around 45 STRs 41. Thus, for HD, a reliable quantification in base space is already very well feasible and will not require signal-level methods such as STRique.

Also, tandem repeats, which are not containing both Guanine and Cytosine bases such as the GAA-tandem repeat linked to Friedreich's Ataxia 1 (FRDA, OMIM: # 229300) are already much better represented than the (G4C2)n-repeat in the current Guppy-flipflop high accuracy model. Increasing for the quantification of GAA-tandem repeats the neural networks window size, already enables consistent quantification results of this sequence.

While we believe that a better model representation for expanded STRs and thus a better quantification of these until now "dark sequences" with general purpose base calling software will become more reliable, we still see a potential use case of STRique specifically for clinical applications of nanopore sequencing.

Here, STRique could serve in conjunction with target enrichment of a larger panel of loci containing potentially expanded STRs and nanopore sequencing as a diagnostic test after genetic counseling and clinical examination of a patient with a family history or clinical presentation of, e.g., a neuropsychiatric disorder. In this scenario, STRique will be employed in a regulated environment.

The key advantages that lead us to favor an HMM approach over deep-learning for this proof of principle study were: the efficient use of training data, the incorporation of prior knowledge into the model and direct access to interpretable model coefficients for evaluation. While not advisable, HMM models theoretically could be derived from a single sequenced read. The strong statistical assumptions of HMMs are justifiable in the case of models trained for a defined stretch of DNA in a defined species and not optimized for performance on arbitrary sequences. On the other hand, there is no need to choose one approach over the other: A HMM can be modified to use deep-learning derived inputs, such as the estimated base-probability tracks from the Flappie/Guppy and HMMs are routinely used in conjecture with deep learning models in other data contexts42. The STRique HMM approach makes it possible to accommodate challenging biological and physical properties of long, repeat-rich DNA even if there is no optimal training data. Once STRique is employed in a clinical environment, the well-understood properties, and greater tractability of Hidden Markov Models in comparison to deep learning models may lower the regulatory hurdles for STRique's and more generally, nanopore sequencing’s clinical adoption.

References

1. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol* **19,** 121 (2018).

2. Sun, J. H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175,** 224–238.e15 (2018).

3. Gymrek, M. A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.* **44,** 9–16 (2017).

4. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).

5. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19,** 286–298 (2018).

6. De Roeck, A. *et al.* Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *bioRxiv* 439026 (2018). doi:10.1101/439026

7. Essebier, A. *et al.* Statistical Enrichment of Epigenetic States Around Triplet Repeats that Can Undergo Expansions. *Front. Neurosci.* **10,** 92 (2016).

8. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet* **6,** 743–755 (2005).

9. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72,** 245–256 (2011).

10. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72,** 257–268 (2011).

11. Beck, J. *et al.* Large C9orf72 hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *Am. J. Hum. Genet.* **92,** 345–353 (2013).

12. van Blitterswijk, M. *et al.* Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol* **12,** 978–988 (2013).

13. Šket, P. *et al.* Characterization of DNA G-quadruplex species forming from C9ORF72 G4C2-expanded repeats associated with amyotrophic lateral sclerosis and frontotemporal lobar degeneration. *Neurobiology of Aging* **36,** 1091–1096 (2015).

14. Xi, Z. *et al.* The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol.* **129,** 715–727 (2015).

15. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65,** 905–914 (1991).

16. Paulson, H. Repeat expansion diseases. *Handbook of Clinical Neurology* **147,** 105–123 (2018).

17. Van Mossevelde, S. *et al.* Clinical Evidence of Disease Anticipation in Families Segregating a C9orf72 Repeat Expansion. *JAMA Neurol* **74,** 445–452 (2017).

18. Song, Q. *et al.* A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. *PLoS ONE* **8,** e81148 (2013).

19. Xi, Z. *et al.* Hypermethylation of the CpG Island Near the G4C2 Repeat in ALS with a C9orf72 Expansion. *The American Journal of Human Genetics* **92,** 981–989 (2013).

20. Russ, J. *et al.* Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. *Acta Neuropathol.* **129,** 39–52 (2015).

21. Cohen-Hadad, Y. *et al.* Marked Differences in C9orf72 Methylation Status and Isoform Expression between C9/ALS Human Embryonic and Induced Pluripotent Stem Cells. *Stem Cell Reports* **0,** (2016).

22. Esanov, R. *et al.* A C9ORF72 BAC mouse model recapitulates key epigenetic perturbations of ALS/FTD. *Mol Neurodegeneration* **12,** 46 (2017).

23. Bell, M. V. *et al.* Physical mapping across the fragile X: Hypermethylation and clinical expression of the fragile X syndrome. *Cell* **64,** 861–866 (1991).

24. Hornstra, L. K., Nelson, D. L., Warren, S. T. & Yang, T. P. High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome. *Hum Mol Genet* **2,** 1659–1665 (1993).

25. Xi, Z. *et al.* Hypermethylation of the CpG-island near the C9orf72 G₄C₂-repeat expansion in FTLD patients. *Hum Mol Genet* **23,** 5630–5637 (2014).

26. Monaghan, K. G., Lyon, E. & Spector, E. B. ACMG Standards and Guidelines for fragile X testing: a revision to the disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics and Genomics. *Genetics in Medicine 2013 15:7* **15,** 575–586 (2013).

27. Lyons, J. I., Kerr, G. R. & Mueller, P. W. Fragile X Syndrome: Scientific Background and Screening Technologies. *The Journal of Molecular Diagnostics* **17,** 463–471 (2015).

28. Cleary, E. M. *et al.* Improved PCR based methods for detecting C9orf72 hexanucleotide repeat expansions. *Molecular and Cellular Probes* **30,** 218–224 (2016).

29. Bahlo, M. *et al.* Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res* **7,** (2018).

30. Almeida, S. *et al.* Modeling key pathological features of frontotemporal dementia with C9ORF72 repeat expansion in iPSC-derived human neurons. *Acta Neuropathol.* **126,** 385–399 (2013).

31. McFarland, K. N. *et al.* SMRT Sequencing of Long Tandem Nucleotide Repeats in SCA10 Reveals Unique Insight of Repeat Expansion Structure. *PLoS ONE* **10,** e0135906 (2015).

32. Ebbert, M. T. W. *et al.* Long-read sequencing across the C9orf72 ‘GGGGCC’ repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegeneration* **13,** 46 (2018).

33. Schüle, B. *et al.* Parkinson’s disease associated with pure ATXN10 repeat expansion. *npj Parkinson's Disease 2017 3:1* **3,** 27 (2017).

34. Ishiura, H. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* **92,** 1 (2018).

35. Tsai, Y.-C. *et al.* Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. *bioRxiv* 203919 (2017). doi:10.1101/203919

36. Mertes, F. *et al.* Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics* **10,** 374–386 (2011).

37. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *bioRxiv* 543439 (2019). doi:10.1101/543439

38. He, F. *et al.* Screening for novel hexanucleotide repeat expansions at ALS- and FTD-associated loci. *Neurol Genet* **2,** e71 (2016).

39. O'Rourke, J. G. *et al.* C9orf72 BAC Transgenic Mice Display Typical Pathologic Features of ALS/FTD. *Neuron* **88,** 892–901 (2015).

40. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the ‘unsequenceable’ genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine 2017 9:1* **9,** 65 (2017).

41. Keum, J. W. *et al.* The HTT CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease. *The American Journal of Human Genetics* **98,** 287–298 (2016).

42. Graves, A., Jaitly, N. & Mohamed, A.-R. Hybrid speech recognition with Deep Bidirectional LSTM. in 273–278 (IEEE). doi:10.1109/ASRU.2013.6707742