

Development of novel methods to quantify somatic CAG repeat expansions in Huntington's Disease

Shota Shibata^{1,2}, Muzhou Wu¹, Sean Koebley³, Andrey Mikheikin^{3,4}, T. Christian Boles⁵, Jason Reed³, Ricardo Mouro Pinto^{1,2} ✉

¹ Center for Genomic Medicine, Massachusetts General Hospital; ² Department of Neurology, Harvard Medical School; ³ Department of Physics and Massey Cancer Center, Virginia Commonwealth University; ⁴ Evizia, Inc., ⁵ Sage Science, Inc.

✉ RMOUROPINTO@MGH.HARVARD.EDU



1. Abstract

Current methodology used to quantify *HTT* CAG repeat expansions suffers from reduced sensitivity and/or inability to detect large expansions. Most methods require PCR amplification of the repeat which is biased towards amplification of shorter alleles. This can result in an underestimation of the actual extent of CAG expansions in a patient's sample. While small-pool PCR followed by Southern blot detection can overcome some of these limitations, this method lacks size resolution, takes days to perform, and has a very small throughput. Next generation short-read sequencing methods can measure a large number of alleles and samples in a single run, as well as provide valuable information in terms of repeat structure (eg. CAA interruptions), but they are unable to generate information on longer alleles (max ~150 CAGs). Capillary-based electrophoresis is the most commonly used assay for its improved resolution, sensitivity and relatively low cost. However, it is also sensitive to PCR bias and has a maximum detection range of ~200 CAGs. Since somatic expansions >1,000 CAGs have been reported in the striatum of HD mutation carriers, as determined by small-pool PCR and Southern blot analysis, this represents a significant shortcoming of existing methods.

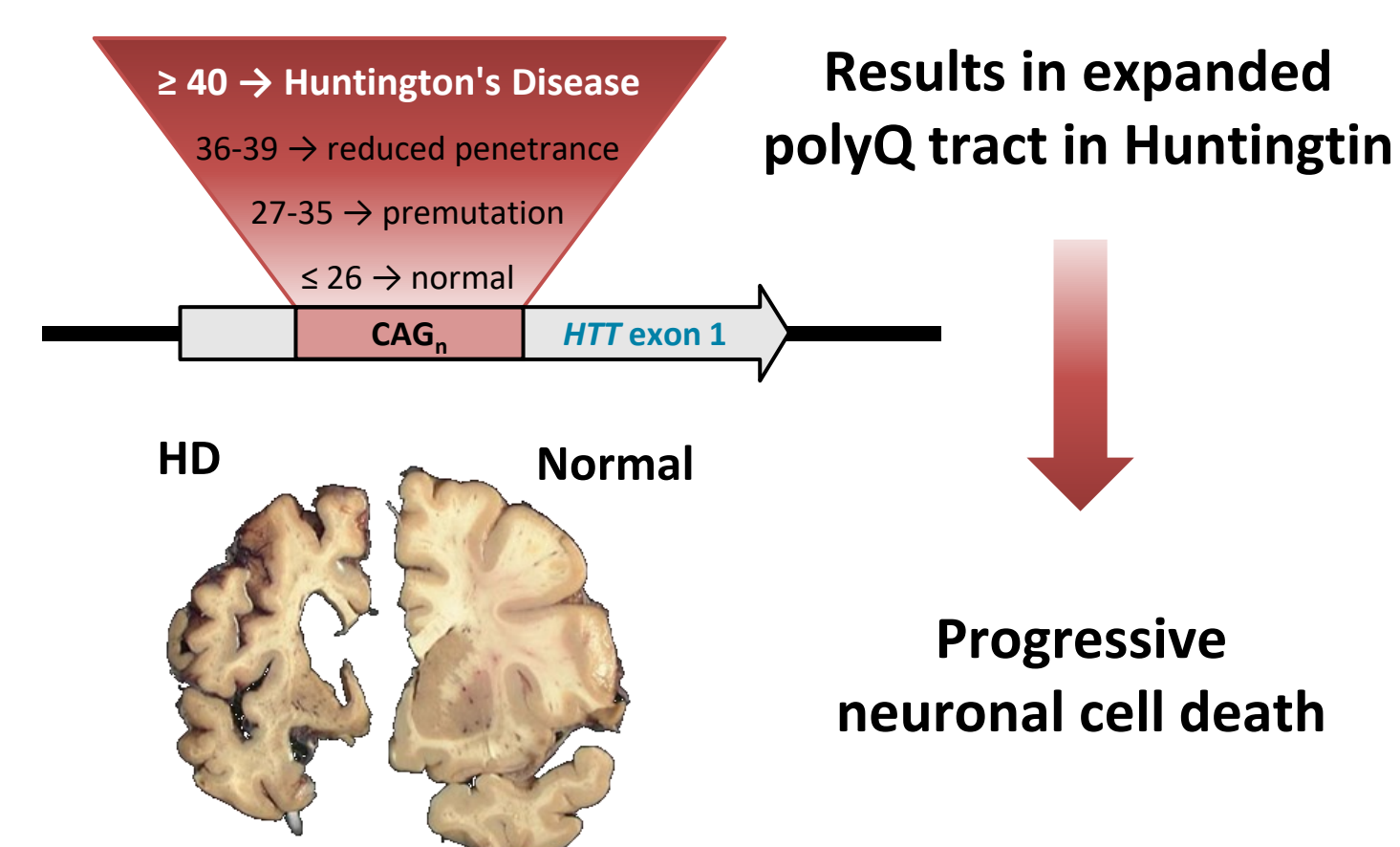
In attempt to address this need, we hereby present preliminary data on the development of two novel methods for *HTT* CAG repeat quantification:

- 1) Single-molecule long-read sequencing: This method combines long-read single-molecule next generation sequencing (Pacbio) with the incorporation of unique molecular barcodes, allowing for deduplication of PCR amplicons and therefore circumvent PCR bias for amplification of shorter alleles. In addition, due to incorporation of unique sample barcodes, this method facilitates the simultaneous quantification of long trinucleotide repeats in multiple samples, from multiple patients at the same time, therefore making for better patient-to-patient or tissue-to-tissue comparisons and substantially reducing costs. Finally, since this is a sequencing-based method, it provides important information on repeat composition and variants such as CAA interruptions, which have been reported as modifiers of HD onset.
- 2) Digital PCR and high-speed atomic force microscopy (HSAFM): this is a single-molecule imaging method that accurately measures the length of amplicons from individual digital PCR reactions, thereby avoiding the PCR bias for shorter alleles. In addition, through automation of the HSAFM measurement process, the method can be scaled to rapidly measure thousands of positive digital PCR reactions.

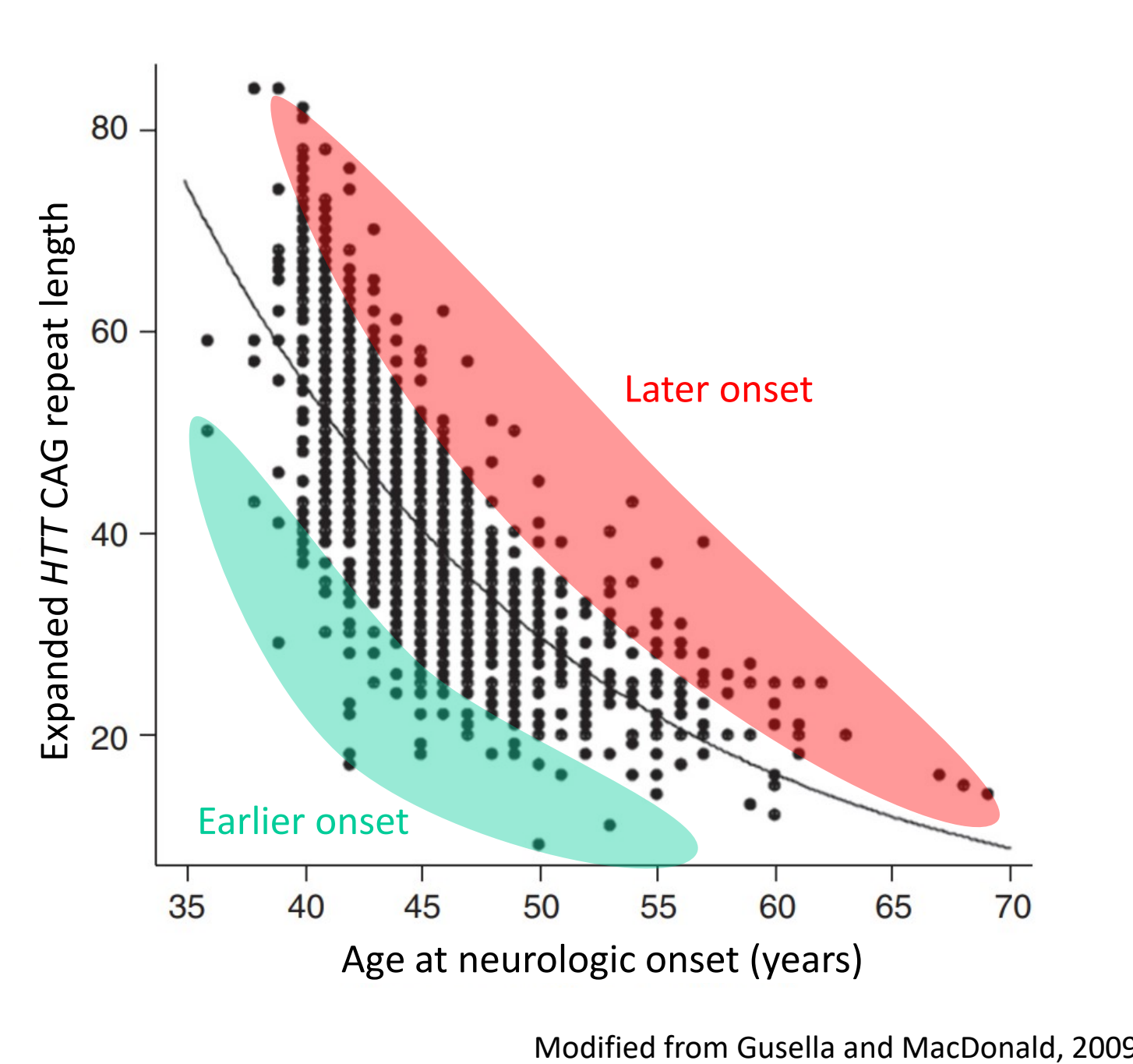
2. Huntington's Disease and Somatic instability

Huntington's Disease (HD)

- ✉ An autosomal dominant disorder caused by the CAG repeat expansion in exon 1 of the *HTT* gene



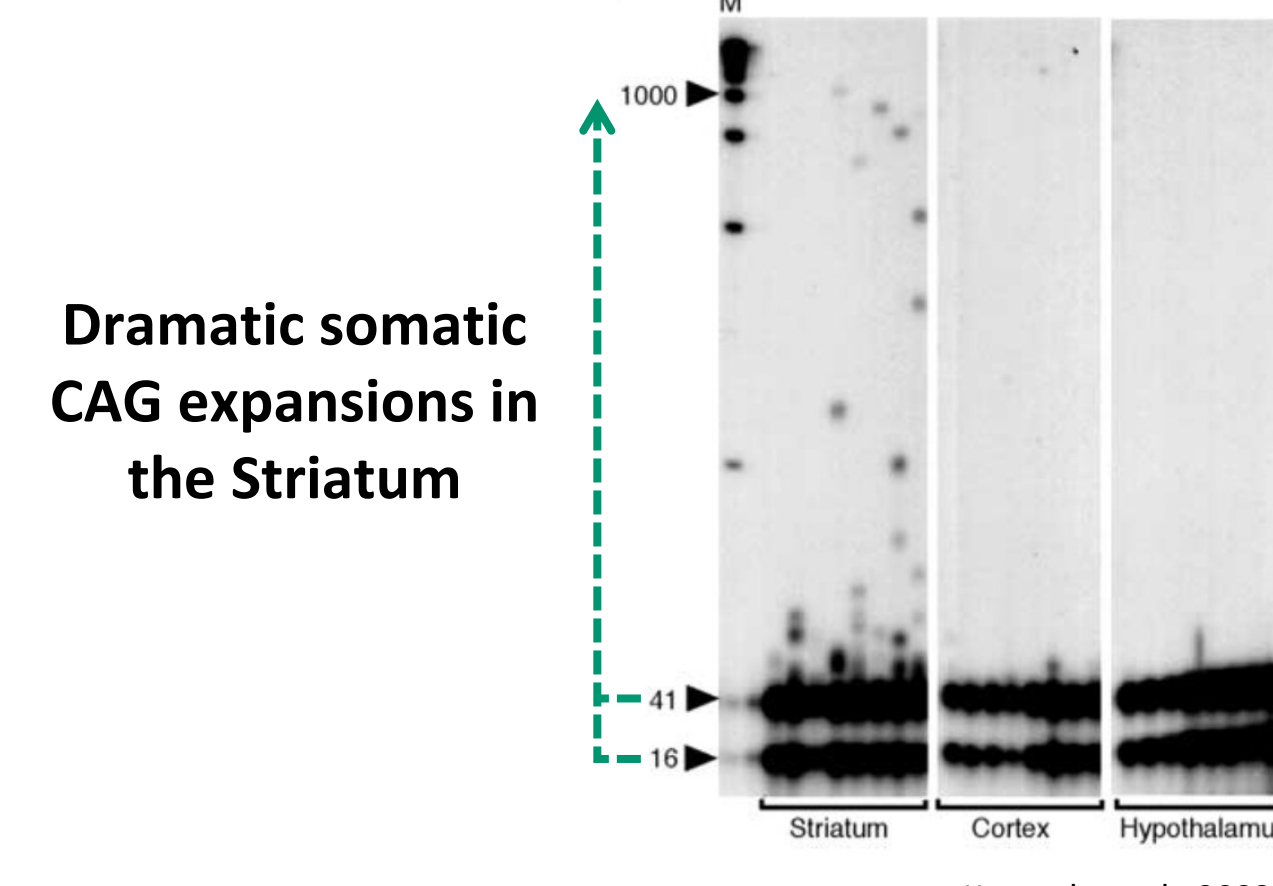
- ✉ Strong inverse correlation is observed between the CAG repeat length and the age at onset



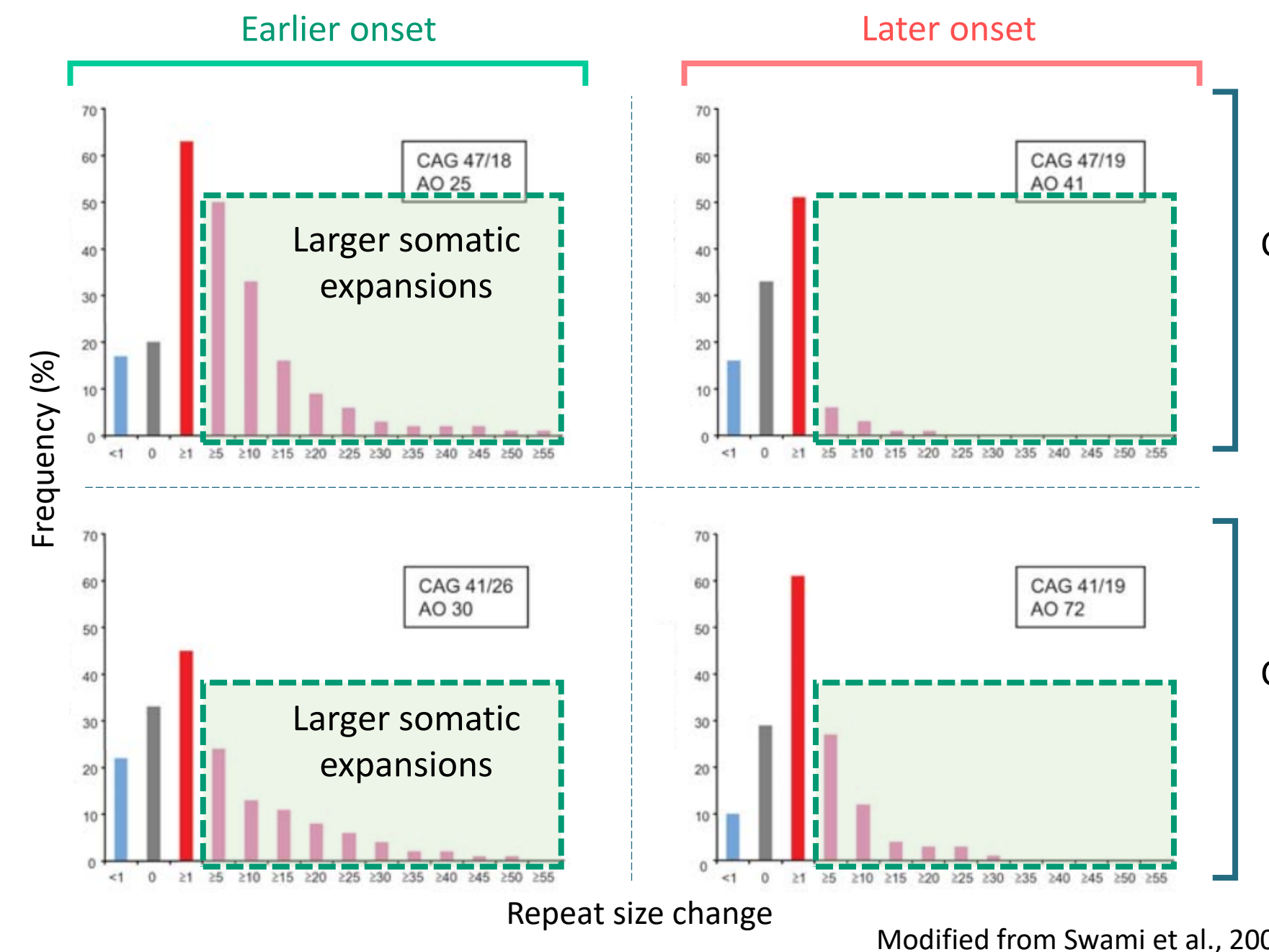
- The CAG repeat expansion is the causative mutation of HD.
- The length of the expanded CAG is a significant determinant of the onset age of HD.

Somatic instability of the expanded CAG in HD

- ✉ The CAG repeats present somatic instability in the postmortem brains of HD patients



- ✉ The somatic CAG expansions in brains are associated with early HD onset



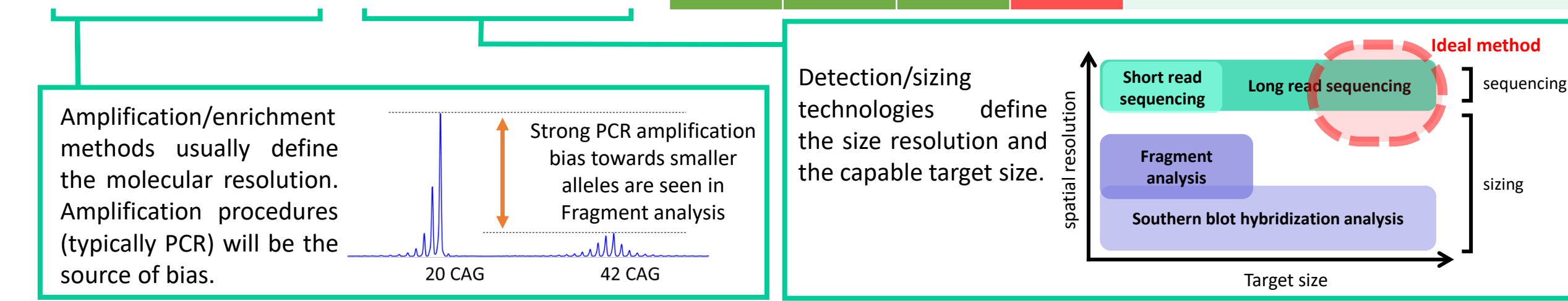
- The somatic instability of the expanded CAG repeats is itself a significant predictor of the onset age of HD.
- This necessitates the evaluation of the expanded repeats at the single-molecule resolution.

Approaches for somatic instability study

- ✉ Current approaches and the call for efficient methods

- Most approaches rely on PCR amplification of the target fragments for the evaluation of the somatic repeat instability. This will be the source of the amplification biases towards the smaller alleles, leading to the underestimation of the repeat sizes.
- The gold standard approach is the small pool PCR (SP-PCR) in combination with Southern blot hybridization analysis. This method mitigate the bias by ensuring the molecular resolution of the starting material, whereas the throughput is low, and it lacks size resolution.
- Currently available alternative approaches are summarized below. The performance of each element was ranked as good (A), fair (B) or poor (C).

Repeat amplification/enrichment	Detection/sizing technology	Single-molecule resolution	Size accuracy	Detection of large repeats	Cost & workload	Summary of Pros & Cons
Small Pool-PCR	Fragment analysis	A	A	B	C	Single-molecule resolution, but increased cost and workload.
	Southern blot hybridization	A	B	A	C	Size evaluation at single-molecule resolution, but lacking size resolution. Time & cost intensive for large sample size.
	Short read sequencing	A	A	C	B	Single-molecule resolution, but still not applicable to large repeats. (~250 CAGs)
Bulk-PCR	Fragment analysis	C	A	B	A	Most commonly used approach. Fast and simple procedures, but large repeats may be affected by amplification bias. Limited detectable fragment size. (~200 CAGs)
	Gel electrophoresis	C	B	A	A	Fast and simple procedures, but lacking molecular & size resolutions.
	Short read sequencing	C	A	C	A	High throughput and scalability, but not applicable to large repeats.
CRISPR/Cas9-based target enrichment	Long read sequencing	C	A	A	B	Ability to size large repeats, but lacking single-molecule resolution. Comparatively low throughput.
	Long read sequencing	A	A	A	C	Single-molecule resolution, ability to size large repeats. Complex target enrichment protocol, reduced yield. (aka "No-Amp")

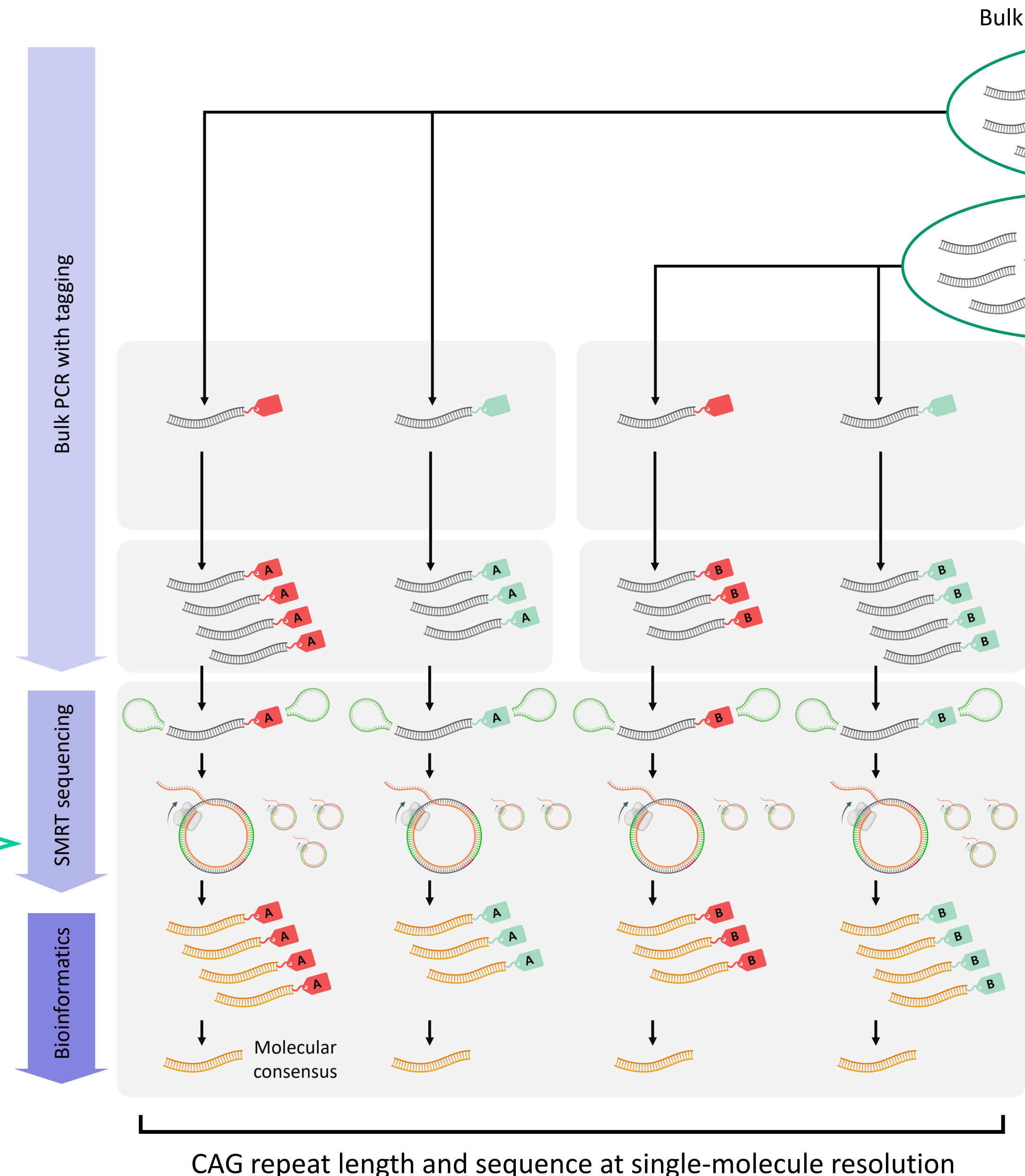
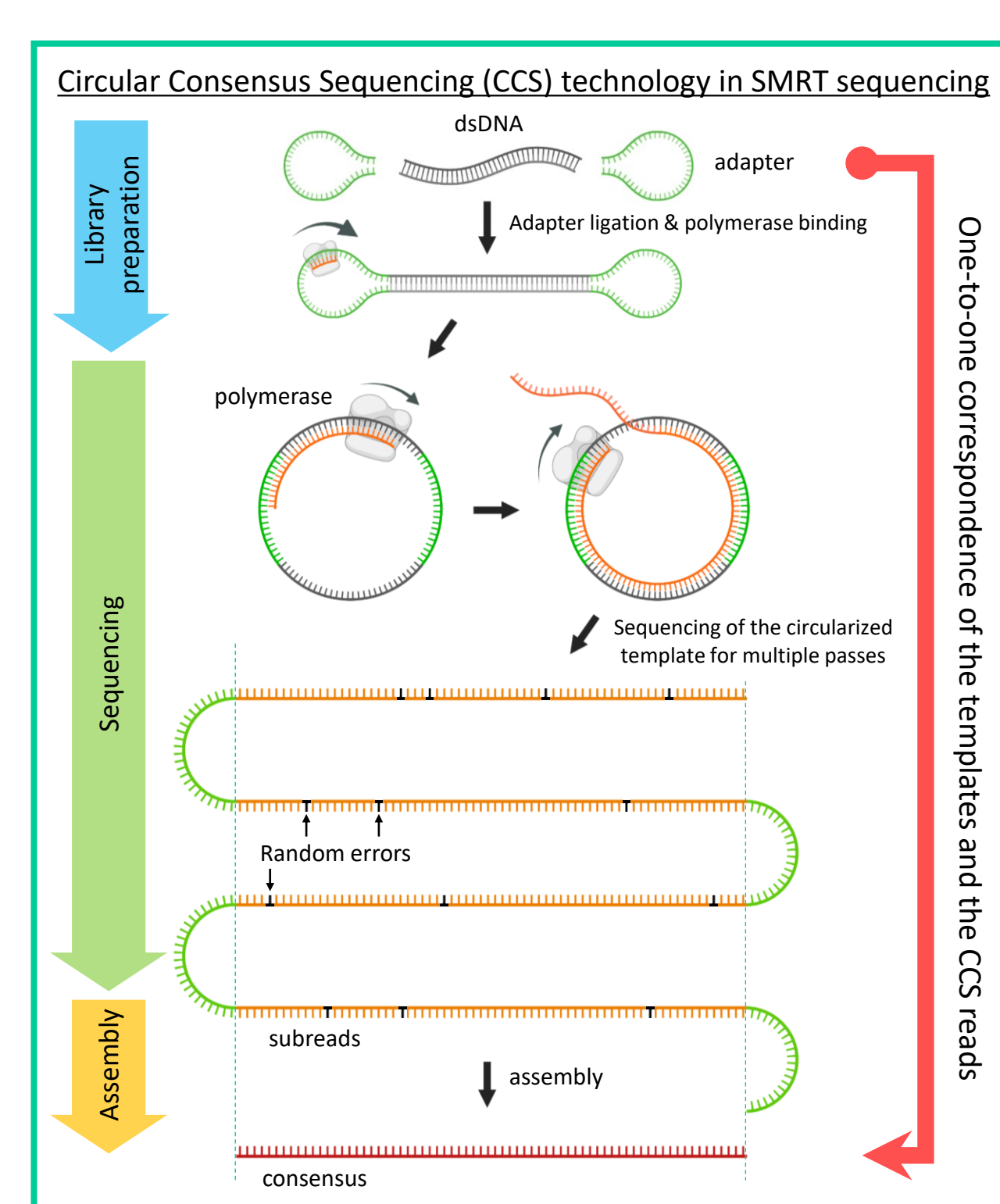


- There is an unmet need for efficient methods of somatic repeat instability characterization.
- Ideally, the method needs: 1) single-molecule resolution, 2) high sizing accuracy, 3) ability to detect large repeat expansions, 4) high throughput, and 5) cost efficiency.

3. Single-molecule long-read sequencing

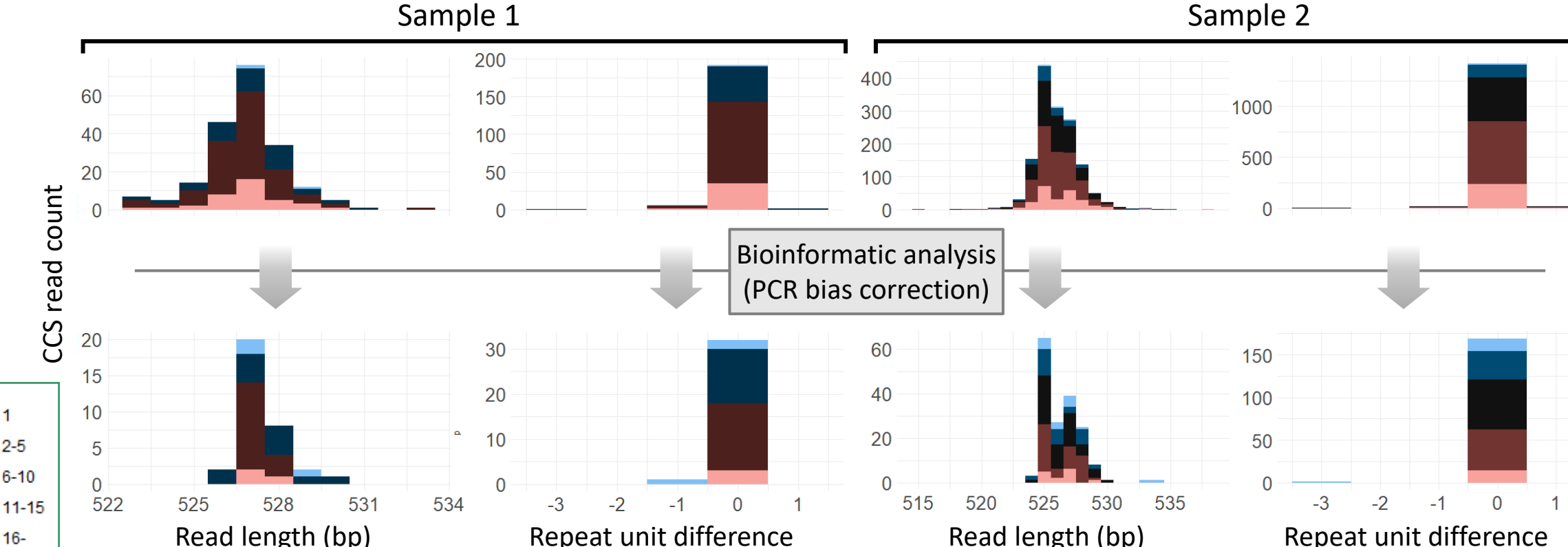
- ✉ The schematic workflow of dual-tagging strategy with single-molecule real-time (SMRT) sequencing

- The dual-tagging strategy utilize the molecular-tags and the sample-tags.
- The molecular consensus sequences are obtained among the reads with the same molecular-tags. This process corrects the PCR biases. The sample-tags enable the multiplexing of samples.
- The one-to-one correspondence of the template dsDNA and the circular consensus sequenced (CCS) is preserved in the process of SMRT sequencing.
- Combining the tagging-strategy and SMRT sequencing, this strategy regain the single-molecule resolution in starting bulk material of multiple samples.



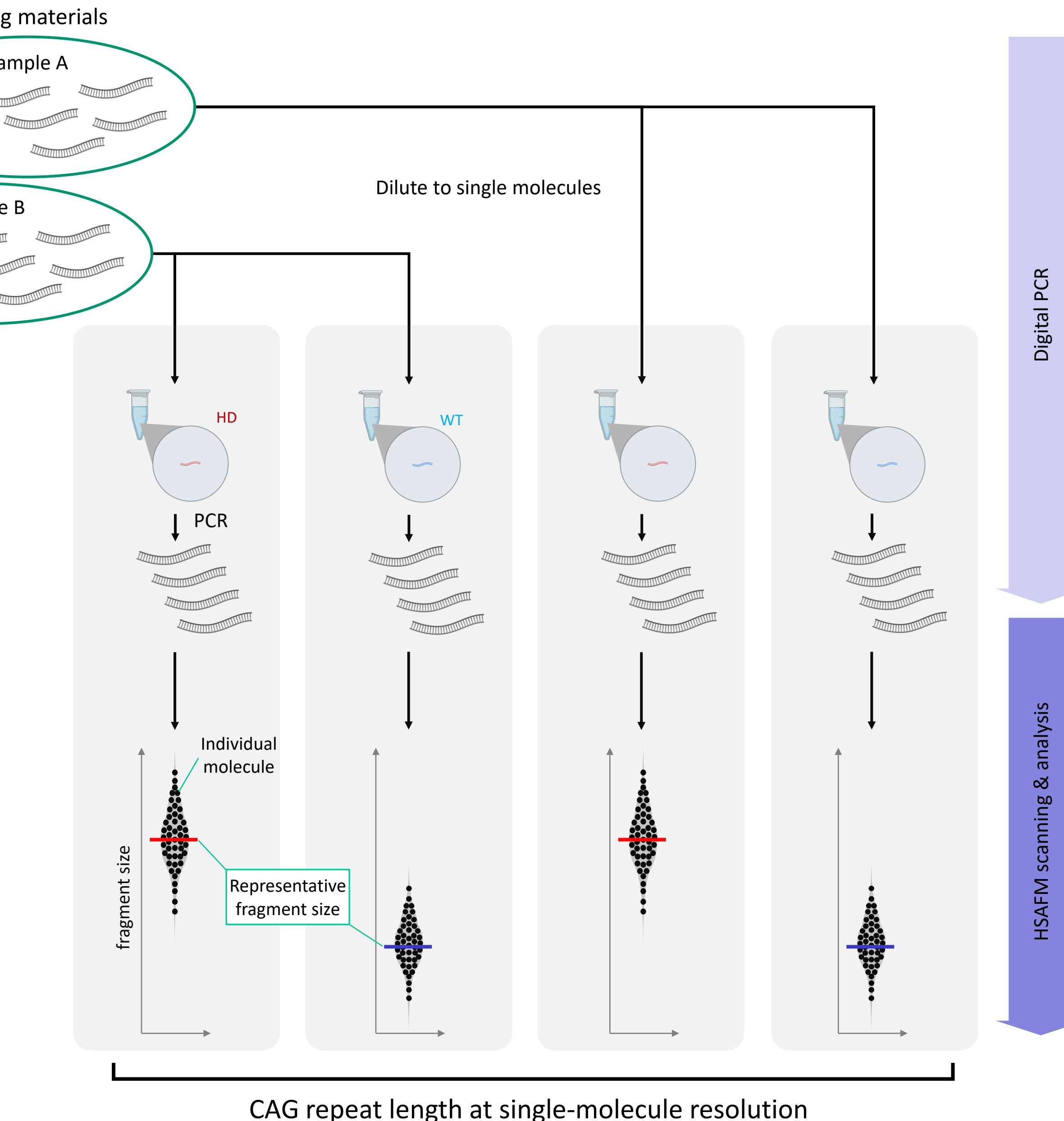
- ✉ Proof of concept (example from non-CAG repeat amplicons)

- The dual-tagging strategy was applied to triplet repeat-containing PCR amplicons from multiple samples.
- The reads assigned to each sample were identified.
- The de-tagging process correcting the PCR bias narrowed down the distributions of the read lengths and the repeat units, resulting in successful calling of repeat lengths at single-molecule resolution.



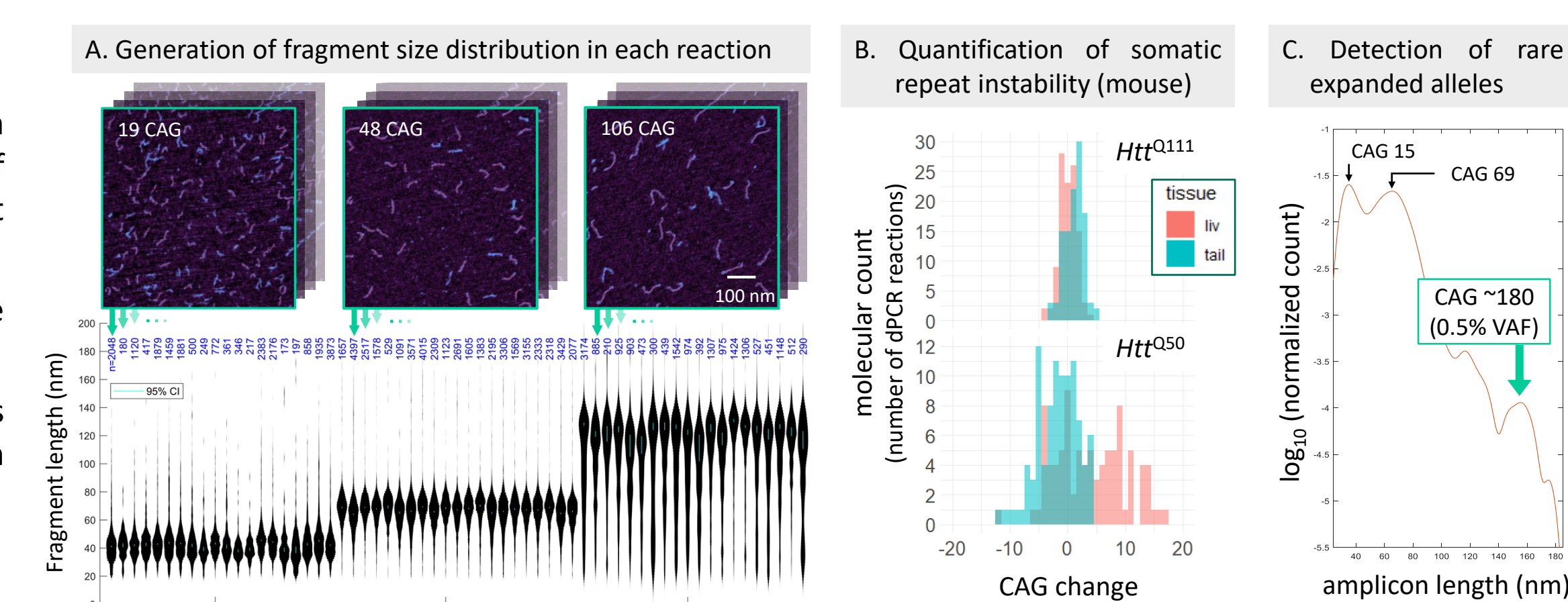
4. Digital PCR & High-Speed Atomic Force Microscopy

- ✉ The schematic workflow of dPCR combined with HSAFM



- ✉ Proof of concept (examples from different datasets)

- The distribution of fragment sizes in each reaction is obtained by automated analysis of HSAFM images. The representative fragment sizes for each reaction are determined (A).
- The repeat units were deduced from the fragment sizes (B).
- The spiked-in rare somatic repeat expansions were successfully detected, proving the high sensitivity of this strategy (C).



5. Conclusion

- ✉ Technological limitations reside in the evaluation of somatic repeat instability despite its striking importance.

- ✉ None of the single applications have sufficient performance for this purpose. Therefore, combination of multiple approaches is necessary to achieve the affordable balance of cost and efficiency.

- ✉ The two novel methods, 1) the long-read sequencing with dual-tagging strategy and 2) the dPCR with HSAFM, were both demonstrated here to be the potential candidates.

- ✉ Efficient quantification measures of somatic mosaicism would contribute to the elucidation of the pathological mechanism and to the development of feasible biomarkers of Huntington's Disease.

Acknowledgement



We are always seeking highly motivated and passionate researchers to join our team! To inquire about open positions and opportunities, please email Dr. Mouro Pinto and include your CV and cover letter.