Comparison of NGS panel and Sanger sequencing for genotyping CAG repeats in the *AR* gene

Maria Santa Rocca[1] (iD) | Margherita Ferrarini[1] | Aichi Msaki[1] | Cinzia Vinanzi[1] |
Marco Ghezzi[1] | Maurizio De Rocco Ponce[1] | Carlo Foresta[1] (iD) | Alberto Ferlin[2]

[1]Unit of Andrology and Reproductive Medicine, Department of Medicine, University of Padua, Padua, Italy

[2]Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy

Correspondence
Carlo Foresta, Unit of Andrology and Reproductive Medicine, Department of Medicine, University of Padua, Via Giustiniani, 2, 35128 Padova, Italy.
Email: carlo.foresta@unipd.it

Abstract

Background: The androgen receptor (AR) is a nuclear receptor, encoded by the *AR* gene on the X chromosome. Within the first exon of the *AR* gene, two short tandem repeats (STR), CAG and GGC, are a source of polymorphism in the population. Therefore, high-throughput methods for screening *AR*, such as next-generation sequencing (NGS), are sought after; however, data generated by NGS are limited by the availability of bioinformatics tools. Here, we evaluated the accuracy of the bioinformatics tool HipSTR in detecting and quantify CAG repeats within the *AR* gene.

Method: The *AR* gene of 228 infertile men was sequenced using NGSgene panel. Data generated were analyzed with HipSTR to detect CAG repeats. The accuracy was compared with the results obtained with Sanger.

Results: We found that HipSTR was more accurate than Sanger in genotyping normal karyotype men (46,XY), however, it was more likely to misidentify homozygote genotypes in men with Klinefelter syndrome (47,XXY).

Conclusion: Our findings show that the bioinformatics tool HipSTR is 100% accurate in detecting and assessing *AR* CAG repeats in infertile men (46,XY) as well as in men with low-level mosaicism.

KEYWORDS

androgen receptor, HipSTR, NGS panel, Sanger, STR

1 | INTRODUCTION

The androgen receptor (AR) (MIM: *313700) belongs to the ligand-activated nuclear receptor superfamily of transcription factors (Davey & Grossmann, 2016).

The main ligands that activate the AR are testosterone and dihydrotestosterone. The AR regulates a plethora of genes important for sex development (Hiort, 2013). The AR gene maps on X chromosome at Xq11.2-12 and the human AR protein consists of 919 amino acids. Genetic variations in this gene have been associated with several pathological conditions such as Androgen Insensitivity Syndrome (MIM: #300068), Kennedy spinal and bulbar muscular atrophy (MIM: #313200), as well as several cancers (Garolla et al., 2005; Giovannelli et al., 2018; Lallous et al., 2016). Given the strong role of the AR in several diseases, a free online database dedicated to all known AR mutations is available since 1994 (http://androgendb.mcgill.ca/)

(Gottlieb, Beitel, Nadarajah, Paliouras, & Trifiro, 2012; Patterson, Hughes, Gottlieb, & Pinsky, 1994). Reported genetic variations in AR include single-point mutations, short insertions/deletions (indels), and splice site mutations (Shukla, Plaga, Shankar, & Gupta, 2016). Another source of polymorphism occurs in two distinct short tandem repeats (STRs) located in exon 1. These trinucleotide repeats, CAG and GGC, are translated into polyglutamine and polyglycine stretches in the transactivation domain of the AR protein (Chamberlain, Driver, & Miesfeld, 1994; Claessens et al., 2008; Need et al., 2009).

These repeats can vary in length and show remarkable ethnic differences (Lund, Tapanainen, Lähdetie, Savontaus, & Aittomäki, 2003). Caucasians have an average of 21–22 CAG repeats and 17–18 GGC repeats (Edwards, Hammond, Jin, Caskey, & Chakraborty, 1992; Ferlin et al., 2004; Platz et al., 1998; Zitzmann & Nieschlag, 2003). It is noteworthy that GGC repeats are generally less polymorphic than CAG repeats (Stanford et al., 1997).

While it is well established that CAG repeats length affects AR transcriptional activity (Tirabassi et al., 2015), it is not known how GGC repeat variations affect AR function (Ferlin et al., 2005). Some studies found that short GGC repeats associated with cancer and male infertility (Ding, Xu, Menon, Reddy, & Barrack, 2005; Ferlin et al., 2004; Sasaki et al., 2005).

Inversely, longer CAG repeats have been associated with male and female infertility (Ashraf, Tariq, & Rehman, 2019; Mobasseri, Babaei, Karimian, & Nikzad, 2018; Xiao et al., 2016). Klinefelter's syndrome is defined by a supernumerary X chromosome (47, XXY) and is the most common genetic cause of male hypogonadism and infertility (Aksglaede et al., 2013; Ferlin et al., 2019; Rocca et al., 2016). Men with this syndrome have various degrees of physiological and intellectual disabilities with the severity depending on the expression of the genetic defect.

Although the extra X chromosome in KS men is inactivated just as it occurs in women, a preferential inactivation of the X chromosome carrying AR allele with longer CAG stretch has been reported in some cases, Suzuki et al. (2001) whereas others have found the contrary to be true (Zitzmann, Depenbusch, Gromoll, and Nieschlag 2004). To date, there is no agreement on which of the two alleles is preferentially expressed.

Women carrying AR allele with long CAG repeats in the active X chromosome show a high risk of developing breast cancer, likely due to a nonrandomly X inactivation (Chen, Wu, Chen, Tsai, & Chien, 2014). Increased frequency of breast cancer is also more common in Klinefelter men (De Sanctis, Fiscina, Soliman, Giovannini, and Yassin 2013). CAG repeats, therefore, have been extensively studied in KS subjects in order to better understand their clinical features (Ferlin et al., 2011; Zinn et al., 2005; Zitzmann et al., 2004).

Overall, the accurate determination of CAG repeats in AR of infertile men, including Klinefelter, should be recommended in clinical practice as it can also predict the risk of developing several tumors (Ferlin et al., 2007; Garolla et al., 2005; Mao et al., 2015).

To date, polymerase chain reaction (PCR) is considered the gold standard method to investigate short tandem repeats and the resulting amplicons are resolved by several molecular technologies. The advancement of sequencing technologies has permitted the fast processing of multiple samples in the detection of single nucleotide variants including the expansion of short tandem repeats. Several analytical methods have evolved in order to determine STR detected by next-generation sequencing (NGS) (Bahlo et al., 2018; Liu, Zhang, Wang, Gu, and Wang, 2017).

Here, we applied HipSTR as a bioinformatics method to assess CAG expansion within AR of 228 men analyzed by NGS and compared it with Sanger.

2 | MATERIAL AND METHODS

2.1 | Subjects

This study was approved by the hospital ethics committee and included 228 subjects retrospectively selected among men referred for fertility evaluation to our Centre (114 men with KS [109 nonmosaic 47,XXY and 5 mosaic 47,XXY/46,XY] and 114 nonsyndromic men [46,XY]). All subjects were of Caucasian ethnicity and Italian origin according to self-report. Men with bone marrow transplant were excluded.

2.2 | Amplification and allele sizing

Genomic DNA was extracted from peripheral blood leukocytes using QIAamp DNA Blood Mini Kit according to the manufacturer's protocol (Qiagen Inc.). The quality of the DNA was examined on a NanoDrop spectrophotometer (Thermo Fisher Scientific Inc).

Determination of the CAG repeat number on *AR* gene was performed by Sanger as previously described (Ferlin et al., 2004). Sanger sequences (Figure 1) were analyzed with the gap4 software of the Staden package (Staden, 1996) available at the UK Human Genome Mapping Project webpage (http://www.hgmp.mrc.ac.uk/).

As an independent method to evaluate CAG repeat length, microsatellite analysis was performed. Briefly, the amplification of CAG polymorphism was performed in a 25-μl PCR mixture containing 50 ng of DNA, 1 mmol/L each primer: 5' end fluorescently labeled with carboxyfluorescein dye (FAM) forward primer 5'-GTGCGCGAAGTGTCCAGAA-3', with its and unlabeled
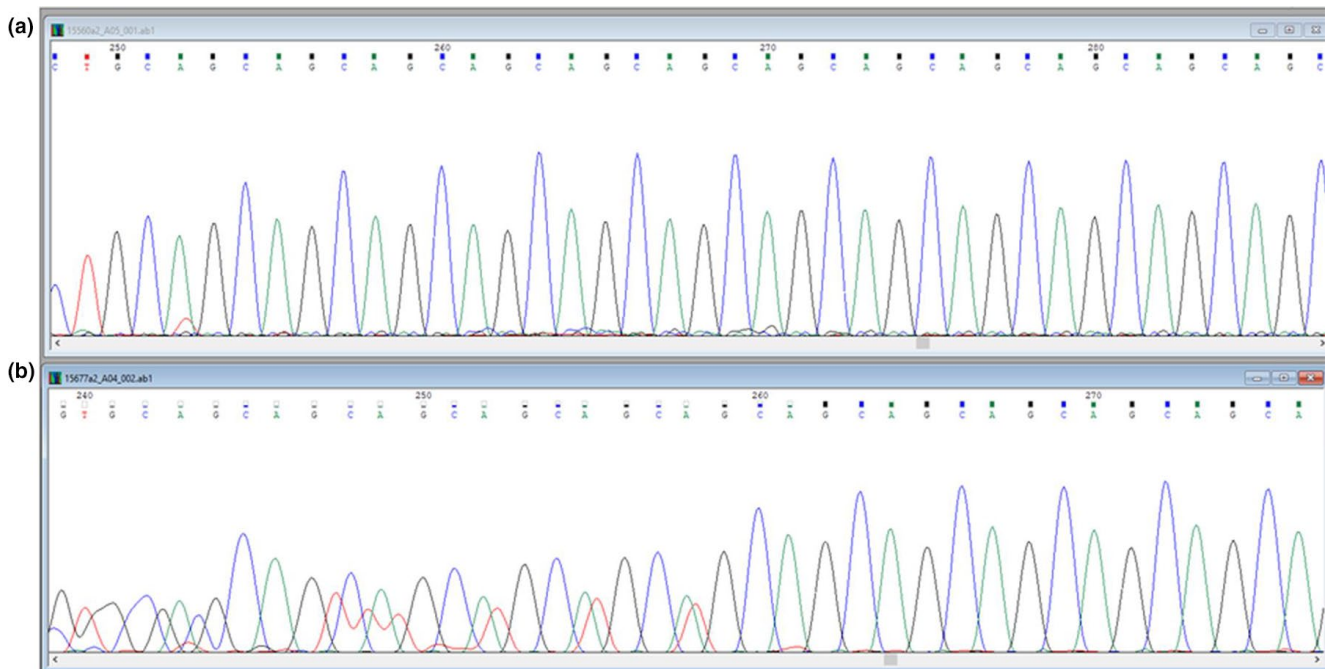
FIGURE 1 Representative Sanger chromatogram of *AR* CAG repeat region in a 46,XY and a Klinefelter 47,XXY men (a) Sanger chromatogram of a male with normal karyotype (46,XY) (b) Sanger chromatogram of a male with Klinefelter syndrome (47,XXY)

reverse primer 5'-TAGCCTGTGGGGCCTCTACG-3' (Ackerman et al., 2012). The PCR mixture contained: PCR buffer, 80 uM dNTPs, 1mM MgCl2, and 1.0 U of Amplitaq Gold (Thermofisher) polymerase. Amplification was performed with an initial denaturation at 95°C for 5 min, followed by 35 cycles of denaturation at 94°C for 1 min, annealing at 53°C for 1 min, extension at 72°C for 1 min, and a final 15-min extension at 72°C. The PCR fragments were resolved by electrophoresis on an automated ABI Prism 3130 XL Genetic Analyzer. GeneScan analysis was performed with PeakScan Software v.1.0 (Applied Biosystems).

2.3 | Sequencing analysis by NGS panel

Sample libraries for sequencing were prepared starting from 80 ng of DNA using AmpliSeq Custom Panel (including probes for AR gene). Genes included in the custom panel were as follows: *AR (OMIM: *313700, NM_000044)* and *TEX11 (OMIM *300311, NM_001003811)* on X chromosome and *FSHR (OMIM: *136435, NM_000145), FSHB (OMIM: *136530, NM_001018080), KLHL10 (OMIM: *608778, NM_152467), NR5A1 (OMIM: *184757, NM_004959), NANOS1 (OMIM: *608226, NM_199461), SEPT12 (OMIM: *611562, NM_144605),* and *SYCP3 (OMIM: *604759, NM_001177948)* on autosomal chromosomes. The libraries were generated using Amplification Library PLUS (24 Reactions) for Illumina according to the manufacturer's protocol (Illumina). The libraries were then loaded on a 500-cycle (2 × 250 paired ends)

reagent cartridge (Illumina) and run on a MiSeq sequencer (Illumina).

For each run, the average depth was of ~100X horizontal coverage to allow for optimal variant calling. BAM files were obtained aligning 250 bp reads to the hg19 reference genome with bwa-mem (v. 0.7.17) (Li & Durbin, 2009) and were sorted and indexed with samtools (v. 0.1.19) (Li et al., 2009). The GGC repeats were not sufficiently covered by design of the probes.

2.4 | STRs genotyping with HipSTR

HipSTR requires a BED file compiled as follows:

chrX	66,765,160	66,765,261	3	34	CAG

The genomic coordinates referred to hg19 were inserted in the first three columns. The motif length was specified in the fourth column. The reference allele length was reported in the fifth column. Finally, in the optional sixth column, the specific analyzed STR locus was named.

In the genome CAG repeats in AR actually go from 66,765,160 to 66,765,225. This is followed by a shorter CAG repeats (6XCAG) 18 base pairs downstream. This poses a problem in the correct alignment of reads in this region. Therefore, the main CAG repeat, the intervening bases, and the shorter CAG repeat were considered as a single STR. Consequently, the reference allele length of the combined STR is 34 (22 + 18/3 + 6).

The hg19 reference sequences of all chromosomes were downloaded from http://hgdownload.cse.ucsc.edu/golde

nPath/hg19/bigZips/chromFa.tar.gz. Chromosome sequences were then concatenated into a single FASTA file and indexed using samtools (v. 0.1.19) (Li et al., 2009).

HipSTR (v. 0.6.2) was used with Mode 1 with the following parameters: --max-str-len 105 and --no-rmdup. This last parameter is necessary when processing PCR-amplified reads.

Samples from 46,XY and KS men were analyzed separately. The option --haploid-chrs chrX was used for 46,XY men because they are haploid for the X chromosome.

HipSTR calls were filtered out if the posterior probability of unphased genotype (Q) was <0.90 and the number of valid reads used for sample's genotype (DP) was <15. The confidence of heterozygous calls was assessed using a combination of the AB and MALLREADS values.

3 | RESULTS

3.1 | 46,XY men

Of 114 men analyzed, 103 samples passed quality filters. We found that genotypes matching between Sanger and HipSTR were 102 and that the only one discordant genotype was correctly called by HipSTR (Table 1). Therefore, the accuracy of Sanger resulted to be lower than HipSTR (102/103 = 0.99 and 103/103 = 1, respectively).

4 | 47,XXY men

Sequencing data from 114 men with 47,XXY karyotype were analyzed. A total of 109 passed quality filters and were further characterized. Capillary electrophoresis determined that 67 were heterozygotes and 42 were homozygotes for _AR_ gene on X chromosome.

TABLE 1 Comparison between Sanger and HipSTR in detecting the genotype of 46,XY men and 47,XXY men

	Results of genotyping
_46,XY (n = 103)_	
Matching genotypes	**102**
Discordant genotypes	**1**
Correct Sanger genotypes	0
Correct HipSTR genotypes	1
47,XXY (n = 109)	
Matching genotypes	**57**
Discordant genotypes	**52**
Correct Sanger genotypes	39
Correct HipSTR genotypes	13

Matching and discordant sequences are in bold.

We found that in 57 cases the Sanger genotype was in agreement with HipSTR. However, there were 52 discordant genotypes and according to capillary electrophoresis. According to microsatellite analysis, Sanger recognized correctly the genotype of 39 individuals, whereas HipSTR only recognized the genotype of 13 individuals (Table 1). Within these 13 cases, Sanger failed to recognize the heterozygosity of four individuals, namely, in the samples ID 3933, 8957, 12587, and 11759 (Figure 2) and the correct number of CAG repeats in nine cases (Table S1).

Therefore, the accuracy of Sanger and HipSTR resulted to be 88% (96/109) and 64% (70/109), respectively.

In Table 2 it is reported the ability of assigning the correct genotype, hence, the probability of identifying true heterozygous or homozygous genotypes. HipSTR identified all heterozygous genotypes, whereas Sanger was limited to 94%. In contrast, HipSTR correctly identified only 7% of the homozygote cases.

5 | DISCUSSION

This is the first study showing the STRs genotyping by HipSTR applied to a target NGS panel and evaluating its accuracy in comparison to Sanger for genotyping _AR_ in 46,XY men and Klinefelter men.

STRs are nucleotide repeats spanning approximately 3% of the whole human genome (Dashnow et al., 2018). As expansions of nucleotide repeats can result in human diseases, the length determination of STR, mapping in coding or regulatory regions, is fundamental for the diagnosis of these pathologies (Paulson, 2018). The pathological STR expansion within _AR_ gene leads to bulbospinal neuronopathy disorder.

In addition to this X-linked neuropathy, the two polymorphic sites of _AR_ gene are intensively studied as risk factors for infertility or cancer. Despite the high risk to miss heterozygosity due to a preferential amplification of one allele (Hamilton et al., 2016), Sanger method is generally used for genotyping _AR_ gene.

Currently, Sanger has been increasingly supplanted by new high-throughput technologies and the increasing progress of NGS has been followed by an equal progress in the bioinformatics field.

Although several bioinformatics tools are available for STRs analysis (Table 3), for this study we chose to use HipSTR (Willems et al., 2017) for the following reasons: (a) it estimates the allele sizes; (b) it has very high accuracy (Bahlo et al., 2018; Gymrek, 2017); (c) it allows a multisample analysis; (d) it analyzes exclusively Illumina data; and (e) it is able to manage differently diploid and haploid genotypes. The latter feature is ideal for genotyping STRs in sex chromosomes.

From the comparison between Sanger and HipSTR analysis, HipSTR resulted to be more accurate than Sanger in genotyping 46,XY men (100% vs. 94%), whereas it resulted less

FIGURE 2 Microsatellite analysis of the *AR* CAG region of four Klinefelter patients in which genotyping was discordant between Sanger and HipSTR

TABLE 2 Sensitivity of Sanger and HipSTR in detecting in 47,XXY subjects true heterozygotes and homozygotes identified by capillary electrophoresis

Method	Heterozygotes (N = 67)	Homozygotes (N = 42)
Sanger	94% (63/67)	100% (42/42)
HipSTR	100% (67/67)	7% (3/42)

efficient than Sanger in genotyping 47,XXY males (64% vs. 88%). Indeed, we found that HipSTR frequently was not able to accurately distinguish homozygous genotypes, calling them as heterozygous genotypes whose two alleles differed for only one triplet. This error-prone situation is likely due to PCR stutter products, however, it can be identified by the STR sizes given in the MALLREADS parameter (Willems et al., 2017).

Nevertheless, in four cases HipSTR prevailed over Sanger in accurately detecting the heterozygous genotypes, whose alleles differ, respectively, for one, two, four, and five triplets (Table S1, Figure 2). In these cases we found MALLREADS and AB values to be informative in clarifying the presence of two true alleles (Table S1).

Specifically, samples with ID 3933, 8957, and 12587 had AB values −0.1, −24.38, and −20.28, respectively, and these results would confirm that true heterozygous calls generally had AB values between 0 and −25. Moreover, for the same samples MALLREADS were indicative of heterozygosity as the most representative alleles differed by more than one triplet and, therefore, they were not consecutive. Interestingly, sample with ID 11759, missed by Sanger and showing a karyotype with 50% of mosaicism 47,XXY/46,XY, was correctly identified by HipSTR, despite it had AB value of −74.83. In this case, MALLREADS highlighted a highly probable second allele (Table S1).

The latter result underscores the risk of missing mild or low-level mosaicism by Sanger. The frequency of mosaic forms is roughly 10%–20% in KS, but it is likely that this prevalence may be higher (Samplaski et al., 2014). Indeed, the broad spectrum of phenotypes in KS could depend on the presence or absence of mosaicism (46,XY/47,XXY) (Tüttelmann & Gromoll, 2010).

While the error of finding a heterozygous genotype with two alleles differing for a single triplet expansion might not

TABLE 3 Tools for genotyping STRs through NGS

Tool	Refs	Estimate STR length	Multisample analysis	Suitable for long STRs
lobSTR	Gymrek, Golan, Rosset, & Erlich (2012)	Y	Y	N
RepeatSeq	Highnam et al. (2013)	Y	N	N
HipSTR	Willems et al. (2017)	Y	Y	N
ExpansionHunter	Dolzhenko et al. (2017)	Y	N	Y
exSTRa	Tankard et al. (2018)	N	Y	Y
STRetch	Dashnow et al. (2018)	Y*	Y	Y

Note: Y* (not tested for PCR + or targeted sequencing data) (ref).

Abbreviations: N, no; NGS, next-generation sequencing; STRs, short tandem repeats; Y, yes.

be clinically relevant, in the case of the possible presence of low level of mosaicism could be worthwhile to further investigate cases that are not detected by traditional methods. In particular, results of HipSTR suggest that preferably 100 metaphases should be analyzed by karyotyping to exclude low level of mosaicisms.

Although Sanger and capillary electrophoresis are still the gold standard methods in detecting STR variation such as the CAG repeats in *AR* gene, the advent of NGS technology represents a big opportunity for investigating massively STR expansions. However, the short reads generated by the Illumina NGS panel are limited to 375 base pairs and, therefore, longer CAG expansions (beyond 30 repeats) could be not covered.

In conclusion, the evaluation of *AR* CAG repeats using bioinformatics tools in men with 47,XXY karyotype must be used with caution. However, we suggest to use NGS panel for the study of *AR* STR in 46,XY infertile men and men with a suspected low-level mosaicism.

CONFLICT OF INTERESTS
The authors declare that there is no conflict of interests.

ORCID
Maria Santa Rocca https://orcid.org/0000-0002-4794-9745
Carlo Foresta https://orcid.org/0000-0002-6576-2183

REFERENCES
Ackerman, C. M., Lowe, L. P., Lee, H., Hayes, M. G., Dyer, A. R., Metzger, B. E., … Urbanek, M. (2012). Ethnic variation in allele distribution of the androgen receptor (AR) (CAG)n repeat. *Journal of Andrology*, *33*(2), 210–215. https://doi.org/10.2164/jandrol.111.013391

Aksglaede, L., Link, K., Giwercman, A., Jørgensen, N., Skakkebæk, N. E., & Juul, A. (2013). 47, XXY Klinefelter syndrome: Clinical characteristics and age-specific recommendations for medical management. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *163C*, 55–63. https://doi.org/10.1002/ajmg.c.31349

Ashraf, M., Tariq, H., & Rehman, R. (2019). Longer trinucleotide repeats of androgen Receptor Gene: Infertility in males. *The Journal of the Pakistan Medical Association*, *69*(3), 446–447.

Bahlo, M., Bennett, M. F., Degorski, P., Tankard, R. M., Delatycki, M. B., & Lockhart, P. J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res*, *7*, 736. https://doi.org/10.12688/f1000research.13980.1

Chamberlain, N. L., Driver, E. D., & Miesfeld, R. L. (1994). The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucleic Acids Research*, *22*(15), 3181–3186. https://doi.org/10.1093/nar/22.15.3181

Chen, H. T., Wu, Y. C., Chen, S. T., Tsai, H. C., & Chien, Y. C. (2014). Androgen receptor CAG repeats, non-random X chromosome inactivation, and loss of heterozygosity at Xq25 in relation to breast cancer risk. *BMC Cancer*, *14*, 144. https://doi.org/10.1186/1471-2407-14-144

Claessens, F., Denayer, S., Van Tilborgh, N., Kerkhofs, S., Helsen, C., & Haelens, A. (2008). Diverse roles of androgen receptor (AR) domains in AR-mediated signaling. *Nuclear Receptor Signaling*, *6*, e008. https://doi.org/10.1621/nrs.06008

Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., … Oshlack, A. (2018). STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology*, *19*(1), 121. https://doi.org/10.1186/s13059-018-1505-2

Davey, R. A., & Grossmann, M. (2016). Androgen receptor structure, function and biology: From bench to bedside. *The Clinical Biochemist. Reviews*, *37*(1), 3–15.

De Sanctis, V., Fiscina, B., Soliman, A., Giovannini, M., & Yassin, M. (2013). Klinefelter syndrome and cancer: From childhood to adulthood. *Pediatric Endocrinology Reviews*, *11*(1), 44–50.

Ding, D., Xu, L., Menon, M., Reddy, G. P., & Barrack, E. R. (2005). Effect of GGC (glycine) repeat length polymorphism in the human androgen receptor on androgen action. *Prostate*, *62*(2), 133–139. https://doi.org/10.1002/pros.20128

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M., van Blitterswijk, M., Narzisi, G., … Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, *27*(11), 1895–1903. https://doi.org/10.1101/gr.225672.117

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., & Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, *12*, 241–253. https://doi.org/10.1016/0888-7543(92)90371-X

Ferlin, A., Bartoloni, L., Rizzo, G., Roverato, A., Garolla, A., & Foresta, C. (2004). Androgen receptor gene CAG and GGC repeat lengths in idiopathic male infertility. *Molecular Human Reproduction*, *10*(6), 417–421. https://doi.org/10.1093/molehr/gah054

Ferlin, A., Dipresa, S., Delbarba, A., Maffezzoni, F., Porcelli, T., Cappelli, C., & Foresta, C. (2019). Contemporary genetics-based diagnostics of male infertility. *Expert Review of Molecular Diagnostics*, *19*(7), 623–633. https://doi.org/10.1080/14737159.2019.1633917

Ferlin, A., Garolla, A., Bettella, A., Bartoloni, L., Vinanzi, C., Roverato, A., & Foresta, C. (2005). Androgen receptor gene CAG and GGC repeat lengths in cryptorchidism. *European Journal of Endocrinology*, *152*(3), 419–425. https://doi.org/10.1530/eje.1.01860

Ferlin, A., Raicu, F., Gatta, V., Zuccarello, D., Palka, G., & Foresta, C. (2007). Male infertility: Role of genetic background. *Reproductive BioMedicine Online*, *14*(6), 734–745. https://doi.org/10.1016/S1472-6483(10)60677-3

Ferlin, A., Schipilliti, M., Vinanzi, C., Garolla, A., Di Mambro, A., Selice, R., … Foresta, C. (2011). Bone mass in subjects with Klinefelter syndrome: Role of testosterone levels and androgen receptor gene CAG polymorphism. *Journal of Clinical Endocrinology and Metabolism*, *96*(4), E739–E745. https://doi.org/10.1210/jc.2010-1878

Garolla, A., Ferlin, A., Vinanzi, C., Roverato, A., Sotti, G., Artibani, W., & Foresta, C. (2005). Molecular analysis of the androgen receptor gene in testicular cancer. *Endocrine-Related Cancer*, *12*(3), 645–655. https://doi.org/10.1677/erc.1.00954

Giovannelli, P., Di Donato, M., Galasso, G., Di Zazzo, E., Bilancio, A., & Migliaccio, A. (2018). The androgen receptor in breast cancer. *Frontiers in Endocrinology*, *9*, 492. https://doi.org/10.3389/fendo.2018.00492

Gottlieb, B., Beitel, L. K., Nadarajah, A., Paliouras, M., & Trifiro, M. (2012). The androgen receptor gene mutations database: 2012 update. *Human Mutation*, *33*(5), 887–894. https://doi.org/10.1002/humu.22046

Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, *44*, 9–16. https://doi.org/10.1016/j.gde.2017.01.012

Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, *22*(6), 1154–1162. https://doi.org/10.1101/gr.135780.111

Hamilton, A., Tétreault, M., Dyment, D. A., Zou, R., Kernohan, K., Geraghty, M. T., … Boycott, K. M. (2016). Concordance between whole-exome sequencing and clinical Sanger sequencing: Implications for patient care. *Molecular Genetics & Genomic Medicine*, *4*(5), 504–512. https://doi.org/10.1002/mgg3.223

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., & Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, *41*(1), e32. https://doi.org/10.1093/nar/gks981

Hiort, O. (2013). The differential role of androgens in early human sex development. *BMC Medicine*, *11*, 152. https://doi.org/10.1186/1741-7015-11-152

Lallous, N., Volik, S. V., Awrey, S., Leblanc, E., Tse, R., Murillo, J., … Cherkasov, A. (2016). Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients. *Genome Biology*, *17*, 10. https://doi.org/10.1186/s13059-015-0864-1

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, Q., Zhang, P., Wang, D., Gu, W., & Wang, K. (2017). Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine*, *9*(1), 65. https://doi.org/10.1186/s13073-017-0456-7

Lund, A., Tapanainen, J. S., Lähdetie, J., Savontaus, M. L., & Aittomäki, K. (2003). Long CAG repeats in the AR gene are not associated with infertility in Finnish males. *Acta Obstetricia et Gynecologica Scandinavica*, *82*(2), 162–166. https://doi.org/10.1034/j.1600-0412.2003.00068.x

Mao, Q., Qiu, M., Dong, G., Xia, W., Zhang, S., Xu, Y., … Jiang, F. (2015). CAG repeat polymorphisms in the androgen receptor and breast cancer risk in women: A meta-analysis of 17 studies. *OncoTargets and Therapy*, *8*, 2111–2120. https://doi.org/10.2147/OTT.S85130

Mobasseri, N., Babaei, F., Karimian, M., & Nikzad, H. (2018). Androgen receptor (AR)-CAG trinucleotide repeat length and idiopathic male infertility: A case-control trial and a meta-analysis. *EXCLI Journal*, *17*, 1167–1179. https://doi.org/10.17179/excli2018-1744

Need, E. F., Scher, H. I., Peters, A. A., Moore, N. L., Cheong, A., Ryan, C. J., … Buchanan, G. (2009). A novel androgen receptor amino terminal region reveals two classes of amino/carboxyl interaction-deficient variants with divergent capacity to activate responsive sites in chromatin. *Endocrinology*, *150*(6), 2674–2682. https://doi.org/10.1210/en.2008-1181

Patterson, M. N., Hughes, I. A., Gottlieb, B., & Pinsky, L. (1994). The androgen receptor gene mutations database. *Nucleic Acids Research*, *22*(17), 3560–3562.

Paulson, H. (2018). Repeat expansion diseases. *Handbook of Clinical Neurology*, *147*, 105–123. https://doi.org/10.1016/B978-0-444-63233-3.00009-9

Platz, E. A., Giovannucci, E., Dahl, D. M., Krithivas, K., Hennekens, C. H., Brown, M., … Kantoff, P. W. (1998). The androgen receptor gene GGN microsatellite and prostate cancer risk. *Cancer Epidemiology, Biomarkers and Prevention*, *7*, 379–384.

Rocca, M. S., Pecile, V., Cleva, L., Speltra, E., Selice, R., Di Mambro, A., … & Ferlin, A. (2016). The Klinefelter syndrome is associated with high recurrence of copy number variations on the X chromosome with a potential role in the clinical phenotype. *Andrology*, *4*(2), 328–334. https://doi.org/10.1111/andr.12146

Samplaski, M. K., Lo, K. C., Grober, E. D., Millar, A., Dimitromanolakis, A., & Jarvi, K. A. (2014). Phenotypic differences in mosaic Klinefelter patients as compared with non-mosaic Klinefelter patients. *Fertility and Sterility*, *101*(4), 950–955. https://doi.org/10.1016/j.fertnstert.2013.12.051

Sasaki, M., Karube, A., Karube, Y., Watari, M., Sakuragi, N., Fujimoto, S., & Dahiya, R. (2005). GGC and StuI polymorphism on the androgen receptor gene in endometrial cancer patients. *Biochemical and Biophysical Research Communications*, *329*(1), 100–104. https://doi.org/10.1016/j.bbrc.2005.01.104

Shukla, G. C., Plaga, A. R., Shankar, E., & Gupta, S. (2016). Androgen receptor-related diseases: What do we know? *Andrology*, *4*(3), 366–381. https://doi.org/10.1111/andr.12167

Staden, R. (1996). The Staden sequence analysis package. *Molecular Biotechnology*, *5*(3), 233–241. https://doi.org/10.1007/BF02900361

Stanford, J. L., Just, J. J., Gibbs, M., Wicklund, K. G., Neal, C. L., Blumenstein, B. A., & Ostrander, E. A. (1997). Polymorphic repeats in the androgen receptor gene: Molecular markers of prostate cancer risk. *Cancer Research*, *57*(6), 1194–1198.

Suzuki, Y., Sasagawa, I., Tateno, T., Ashida, J., Nakada, T., Muroya, K., & Ogata, T. (2001). Mutation screening and CAG repeat length analysis of the androgen receptor gene in Klinefelter's syndrome patients with and without spermatogenesis. *Human Reproduction*, *16*(8), 1653–1656. https://doi.org/10.1093/humrep/16.8.1653

Tankard, R. M., Bennett, M. F., Degorski, P., Delatycki, M. B., Lockhart, P. J., & Bahlo, M. (2018). Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *American Journal of Human Genetics*, *103*(6), 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015

Tirabassi, G., Cignarelli, A., Perrini, S., Delli Muti, N., Furlani, G., Gallo, M., … Balercia, G. (2015). Influence of CAG repeat polymorphism on the targets of testosterone action. *International Journal of Endocrinology*, *2015*, 298107. https://doi.org/10.1155/2015/298107

Tüttelmann, F., & Gromoll, J. (2010). Novel genetic aspects of Klinefelter's syndrome. *Molecular Human Reproduction*, *16*, 386–395. https://doi.org/10.1093/molehr/gaq019

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., & Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, *14*(6), 590–592. https://doi.org/10.1038/nmeth.4267

Xiao, F., Lan, A., Lin, Z., Song, J., Zhang, Y., Li, J., … Yang, X. (2016). Impact of CAG repeat length in the androgen receptor gene on male infertility - a meta-analysis. *Reproductive Biomedicine Online*, *33*(1), 39–49. https://doi.org/10.1016/j.rbmo.2016.03.012

Zinn, A. R., Ramos, P., Elder, F. F., Kowal, K., Samango-Sprouse, C., & Ross, J. L. (2005). Androgen receptor CAGn repeat length influences phenotype of 47, XXY (Klinefelter) syndrome. *Journal of Clinical Endocrinology and Metabolism*, *90*(9), 5041–5046. https://doi.org/10.1210/jc.2005-0432

Zitzmann, M., Depenbusch, M., Gromoll, J., & Nieschlag, E. (2004). X-chromosome inactivation patterns and androgen receptor functionality influence phenotype and social characteristics as well as pharmacogenetics of testosterone therapy in Klinefelter patients. *Journal of Clinical Endocrinology and Metabolism*, *89*(12), 6208–6217. https://doi.org/10.1210/jc.2004-1424

Zitzmann, M., & Nieschlag, E. (2003). The CAG repeat polymorphism within the androgen receptor gene and maleness. *International Journal of Andrology*, *26*(2), 76–83. https://doi.org/10.1046/j.1365-2605.2003.00393.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.