Hum Genet (2017) 136:591–603
DOI 10.1007/s00439-017-1788-5

REVIEW

Human Y chromosome copy number variation in the next generation sequencing era and beyond

Andrea Massaia[1,2] · **Yali Xue**[2]

Received: 14 February 2017 / Accepted: 25 March 2017 / Published online: 4 April 2017

Abstract The human Y chromosome provides a fertile ground for structural rearrangements owing to its haploidy and high content of repeated sequences. The methodologies used for copy number variation (CNV) studies have developed over the years. Low-throughput techniques based on direct observation of rearrangements were developed early on, and are still used, often to complement array-based or sequencing approaches which have limited power in regions with high repeat content and specifically in the presence of long, identical repeats, such as those found in human sex chromosomes. Some specific rearrangements have been investigated for decades; because of their effects on fertility, or their outstanding evolutionary features, the interest in these has not diminished. However, following the flourishing of large-scale genomics, several studies have investigated CNVs across the whole chromosome. These studies sometimes employ data generated within large genomic projects such as the DDD study or the 1000 Genomes Project, and often survey large samples of healthy individuals without any prior selection. Novel technologies based on sequencing long molecules and combinations of technologies, promise to stimulate the study of Y-CNVs in the immediate future.

✉ Andrea Massaia
a.massaia@imperial.ac.uk

✉ Yali Xue
ylx@sanger.ac.uk

[1] National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK

[2] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK

Introduction

The Y chromosome, here referring to its haploid, male-specific portion (MSY), is a unique segment of the human genome. It is non-essential for the life of an individual but required for male sexual differentiation, and evidence for its role in human biology beyond male reproduction is growing (Bellott et al. 2014). Its functional uniqueness is matched by its structural complexity: the human Y is rich in repeated elements and segmental duplications, which cover ~35% of its length (Skaletsky et al. 2003). While polymorphisms for presence and absence of repeated elements are common in the rest of the genome (Conrad et al. 2010; Mills et al. 2011; Sudmant et al. 2015), only two such polymorphisms, the insertion of an *Alu* element, that in the phylogeny identifies haplogroup DE (Hammer 1994), and the insertion of a LINE-1 element in a subgroup of haplogroup O (Santos et al. 2000) are known in the Y chromosome. Nevertheless, repeated elements are tied to other classes of genomic rearrangements: they are believed to be directly involved in one of the proposed mechanisms for structural rearrangements (non-allelic homologous recombination, NAHR) and their frequent presence near putative CNV breakpoints has been described in the Y chromosome (Poznik et al. 2016) (Fig. 1), as in the rest of the genome (Conrad et al. 2010). Intuitively, the abundance of repeats is a possible cause (Redon et al. 2006), but also a plausible consequence of frequent structural rearrangements. For instance, the palindromes in ampliconic regions (Skaletsky et al. 2003) show a high arm-to-arm sequence similarity, which is proposed to be maintained by frequent gene conversion events (Rozen et al. 2003): this may have the effect of preserving important, fertility-related genes from decay over evolutionary timescales by both reducing the

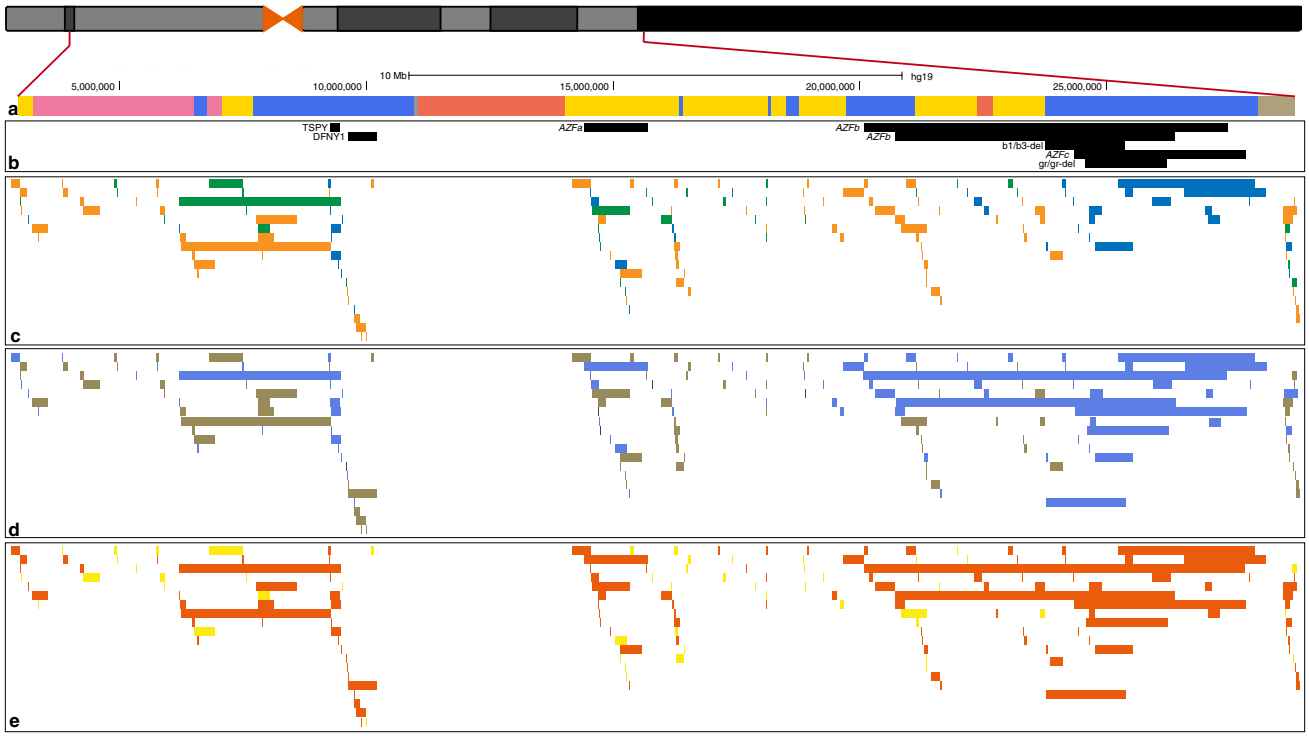


Fig. 1 Summary of CNVs on the human Y chromosome. **a** Male-specific euchromatic region of the Y chromosome. The Y-specific unique region is shown in *yellow*, the X–Y transposed region in *red*, Y-specific repeats in *blue*, heterochromatic segments in *purple* and other regions in *grey*. **b** Medically important Y-CNVs. **c** CNVs discovered from population studies. Deletions are shown in *orange*, duplications in *green*, and deletions/duplications in *blue*. **d** Y-CNV mutation events inferred from the available data. Single events are shown in *yellow*, recurrent events in *blue* and unknown ones in *dark grey*. **e** Y-CNVs associated with segmental duplications or other repeats are shown in *dark orange*, and non-repeat-associated ones in *yellow*

accumulation of deleterious mutations when coupled with purifying selection, and also by facilitating the fixation of potential beneficial mutations when coupled with positive selection (Betran et al. 2012).

While repetitive sequences may facilitate structural rearrangements, they also make their detection harder: emblematically, when the sequence of the Y chromosome was published in 2003, *Nature*'s cover described it as "a genetic hall of mirrors". Most current detection methods are tailored to the diploid genome, and their prior expectations about copy number may not be adequate to the haploidy of the Y chromosome. Long, highly similar inter- and intra-chromosomal multicopy sequences make reference-based methods unreliable, making it difficult to map sequencing or intensity data correctly, and to univocally assign observed variation to a specific region, an effect defined as "shadowing" (Wei et al. 2015) ("Box 1" and Fig. 2). Despite these difficulties, several regions of the chromosome are well known to be prone to specific rearrangements (Jobling 2008), and these have continued to be investigated by focused studies in the past few years. The abundance of information about these regions mostly depends on some specific features, which historically led to the discovery of these variants, such as the effects on male fertility of azoospermia factor (*AZF)* loci (Vogt et al. 1996), the high and hypervariable copy number of the *TSPY* gene (Tyler-Smith et al. 1988), or the failure in sex testing caused by *AMELY* deletions (Santos et al. 1998). Wider studies of Y-CNVs have been scarce until recent years, and genome-wide CNVs investigations touched the Y chromosome only marginally. The pioneering study by Redon et al. (2006), which employed a combination of BAC arrays and comparative genomic hybridization to build the first CNV map of the human genome, reported over 250 variants on the Y chromosome. What looked like a promising start for Y-CNVs turned into a notable exception as subsequent genome-wide CNVs studies largely ignored the Y chromosome, either because they were carried out in females (Conrad et al. 2010) or simply because they reported a very small number of Y-CNVs, if any (Mills et al. 2011; The 1000 Genomes Project Consortium 2012). In the 1000 Genomes Project Phase 3, only six structural variants on the Y chromosome were described by genome-wide analyses (Sudmant et al. 2015; The 1000 Genomes Project Consortium 2015). In the latest structural variation (SV) data release by the Genomes of the Netherlands Project (Francioli et al. 2015; The Genome of the

Netherlands Consortium 2014), which also produced a dedicated study of structural variants (Hehir-Kwa et al. 2016), only 4556 out of 1,851,571 structural variants (less than 0.25%) were mapped to the Y chromosome. Such studies focused on the diploid genome; it is only recently that similar high-throughput, unbiased studies have focused on the Y chromosome.

In this review, we first discuss the methodologies used in the past and present and some potential developments in the future for studying Y-CNVs, then the observations we have gained so far from the targeted studies, and finally the chromosome-wide studies.

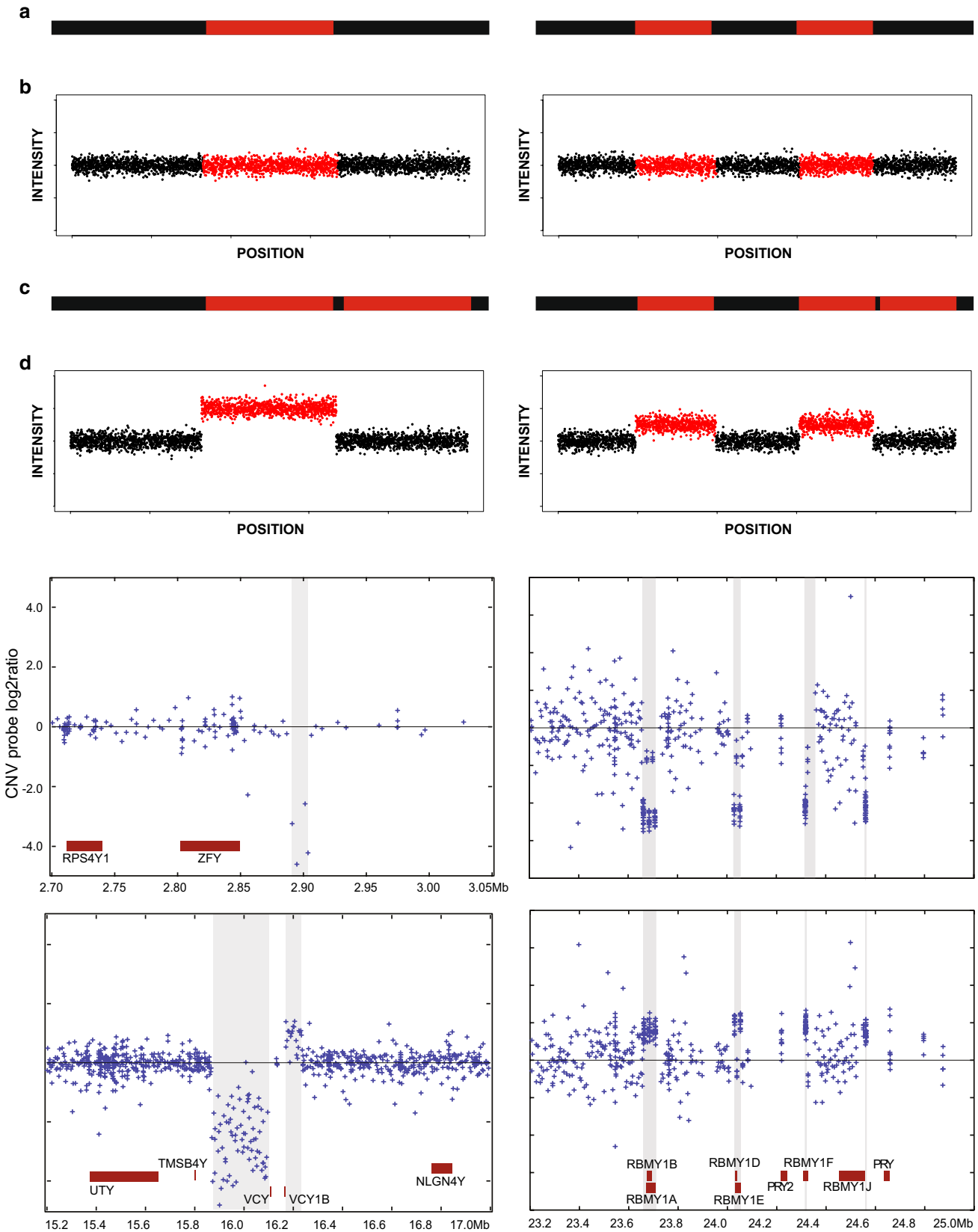
Methodologies for CNV studies

Several methods have been employed in Y-CNVs studies, and the choice is mostly driven by the resolution and power offered by each method, and the intrinsic features of the information produced. A summary of the methods presented below is given in Table 1.

Cytogenetics allows direct observation of the alternative structures generated by structural rearrangements. However, cytogenetic methods have low throughput due to the laboriousness of the technology that often requires cell culture and high levels of skill in implementation and interpretation, making it inconvenient to analyse more than a handful of samples at once; the resolution falls in a wide range depending on the specific method. Karyotyping allows the detection of aneuploidies and rearrangements down to ~5 Mb in size; it has been used to observe Y aneuploidies ranging from zero copies, as in Turner syndrome (Legro 2012), up to four copies (Paoloni-Giacobino and Lespinasse 2007), and provides a validation method to confirm mosaicism (Poznik et al. 2016). It is notable that the first evidence for an *AZF* locus on the Y (Tiepolo and Zuffardi 1976) and age-related somatic loss of Y (Jacobs et al. 1963) were from cytogenetic analyses. Higher resolution can be achieved with fluorescence in situ hybridization (FISH) techniques, which can be performed on interphase nuclei or metaphase chromosomes, in a setup similar to non-fluorescent karyotyping, but also on linearized chromatin fibres. Fibre-FISH allows for greater detail, ranging from ~100 kb when using BAC-clone derived probes, down to a few kilobases when using custom probes; moreover, using multiple probes can reveal inversions, a class of rearrangements otherwise difficult to detect. FISH-based methods can be used as validation in large studies (Poznik et al. 2016), but can also be used as the main investigation method when targeting specific variants, such as the change in size of the long arm heterochromatin block (Repping et al. 2003, 2006).

PCR-based methods are also, generally speaking, low-throughput, as the PCR technology produces short-range information (with a single reaction usually limited to less than 10 kb) and is then impractical to employ alone in genome-wide or chromosome-wide studies. At the same time, however, PCR is easily scalable to a large number of reactions: a recently described application, droplet digital PCR (Hindson et al. 2011) can process up to 2 million reactions in a single workflow. PCR approaches are then ideal in the screening of large cohorts for a limited number of specific variants. For instance, the study by Rozen et al. (2012) used PCR-based STS detection to assess the *AZFc* structure in over ~20,000 individuals. PCR also allows multiplexing, enabling the analysis of more than one region at the same time: this has been exploited to design the *AMELX/Y* sex test (Sullivan et al. 1993), but also to design clinical tests for *AZF* microdeletions (Vogt and Bender 2013). Simultaneous reactions can also be employed to directly count the number of members in a gene family by real-time quantitative PCR (Kumari et al. 2012). Real-time PCR has also been used to design a test for the Y chromosome in free foetal DNA in maternal blood (Boon et al. 2007), thanks to its high sensitivity and specificity. Above all, Sanger sequencing of PCR products makes it possible to reach base-pair resolution. This allows, for instance, targeting and validation of breakpoints and inference of mutational mechanism from the surrounding sequence. Together with the high scalability, this makes PCR a gold standard validation method, even in large-scale studies (Mills et al. 2011).

Microarray technologies infer CNVs by interpreting intensity signals, rather than detecting them directly. Compared to cytogenetics and PCR-based methods, microarray-based methods can produce a notably higher amount of information. For instance, the Illumina Infinium Core-24 Kit analyses up to ~600,000 markers, promising a throughput of 2800 samples per week. Different technologies exist, with resolution from ~100 kb in BAC-clone based arrays (Redon et al. 2006), to ~500 bp with high-resolution oligonucleotide probes (Conrad et al. 2010), including SNP arrays. Array-based methods have been used in many large-scale CNV studies, including both genome-wide and Y-specific studies, either as the main data source (Conrad et al. 2010; Johansson et al. 2015; Redon et al. 2006; Wei et al. 2015), or to validate discoveries from sequencing (Mills et al. 2011; Poznik et al. 2016). Besides the advantages of high data output and easy scalability, however, microarrays present a critical limit, in that they are based on sequence similarity between probe and target. This feature is especially problematic in the presence of long, nearly identical repeats, such as those on sex chromosomes. In these instances, array-based methods (and specifically, technologies based on shorter probes such as SNP arrays and array CGH) will not be able to assign an unequivocal signal to each of the repeated units; a change in copy number at one


a
b
INTENSITY
POSITION
c
d
CNV probe log2ratio
4.0
2.0
0
-2.0
-4.0
RPS4Y1
ZFY
2.70
2.75
2.80
2.85
2.90
2.95
3.00
3.05Mb
TMSB4Y
UTY
VCY
VCY1B
NLGN4Y
15.2
15.4
15.6
15.8
16.0
16.2
16.4
16.6
16.8
17.0Mb
RBMY1B
RBMY1A
RBMY1D
RBMY1E
RBMY1F
PRY2
RBMY1J
PRY
23.2
23.4
23.6
23.8
24.0
24.2
24.4
24.6
24.8
25.0Mb

◂**Fig. 2** Shadowing effect for intensity data. The *top half* shows schematic representations of CNVs and the corresponding intensity data plots. **a** A unique region (*left*) or duplicated region (*right*) in the reference genome is shown in *red*. **b** Corresponding *plots* showing the intensity signal for each probe, here represented by a single *dot*, on the Y axis, and the position for the probe on the X axis. **c** A hypothetical duplication of the unique region (*left*), and of one of the copies of the duplicated region (*right*). **d** The unique region will show a stronger increase in signal (*left*), as compared to the duplicated region (*right*); in the duplicated region, moreover, the increase will be detected in both reference copies, as the method is unable to distinguish between them. The *bottom half* shows real examples for both CNVs in unique regions (*left*) and in a repeated region showing the shadowing effect (*right*) (from Wei et al. 2015). On the *right*, the *RBMY* gene copies all show co-ordinated intensity changes

repeat will be reported as a much smaller change, of the same sign (increase or decrease), at each one of the repeats, making it impossible to tell which one is actually mutating. ("Box 1" and Fig. 2).

Next generation sequencing (NGS) is now established as the prime data-generation method in genomics, and Y chromosome CNV analysis is no exception. NGS offers high throughput comparable to, and even higher than, microarray-based methods; it can potentially achieve base-pair resolution, and indeed has been employed recently as the main data source to study CNVs in the Y chromosome (Espinosa et al. 2015; Poznik et al. 2016) as well as in the rest of the genome (Hehir-Kwa et al. 2016; Mills et al. 2011; Sudmant et al. 2015). It should be noted, however, that the Y chromosome genomic context amplifies NGS's intrinsic limitations. First, sequencing data analyses require mapping to a reference sequence, a step which is confounded by the highly repetitive nature of the Y (Jobling 2008). Uncertainty in mapping produces the "shadowing" effect mentioned earlier and shifts the focus of data analyses, explaining the abundance of computational methods developed to handle NGS data in CNV studies (Pirooznia et al. 2015); methods which are, however, usually tailored to the diploid genome, and may require additional care when applied to the MSY. Second, while sequencing can theoretically reach base-pair resolution, technical limitations such as low depth (median read depth of 4.3×) can preclude the identification of smaller variants; for instance, the smallest CNV identified by Poznik et al. (2016) on the Y chromosome was 2.5 kb. Furthermore, most NGS platforms rely on short reads, thus producing short-range information that can fail to detect complex rearrangements, including copy neutral events such as inversions and translocations, and produces limited information about breakpoints. In this respect, the development of long read sequencing technologies such as PacBio (Rhoads and Au 2015) or Oxford Nanopore (Laver et al. 2015) appears promising. Long-range information is also produced by 10X Genomics, through short read sequencing of individually barcoded long molecules, or "linked-read sequencing" (Zheng et al. 2016). The preprint by Spies et al. (2016) showed how this method can be used to resolve complex structural variants. The same group published a similar study using a similar technology of "synthetic long reads" developed by Illumina, named TruSeq (Bishara et al. 2015). These

Table 1 Summary of CNV detection and follow-up methods used currently or in the past in Y chromosome studies

Method	Resolution*	Throughput	Analysis procedure	Application
Karyotyping	5 Mb	Low	Visual inspection	Genome-wide detection
Interphase FISH	100 kb–1 Mb	Low	Visual inspection	Validation, detection of complex structures
Fibre-FISH	10 kb–1 Mb	Low	Visual inspection	Validation, detection of complex structures
BAC array CGH	100 kb	High	Intensity detection and processing	Genome-wide detection
Oligo array CGH	500 bp	High	Intensity detection and processing	Genome-wide detection, validation
Short read WG sequencing	1 bp	High	Read mapping and variant calling, de novo assembly	Genome-wide detection
Long read WG sequencing	1 bp	High	Read mapping and variant calling, de novo assembly	Genome-wide detection
qPCR, paralogue ratio test (PRT)	200 bp	Medium	Fluorescence detection	Validation, copy number quantification
Digital droplet PCR (ddPCR)	200 bp	High	Fluorescence detection	Validation, copy number quantification
Sequence-tagged site (STS)	200 bp	Medium	Electrophoresis: band presence/absence	Validation, targeted assay
Breakpoint PCR	1 bp	Low	Electrophoresis: band presence/absence	Validation and refinement

* Resolution indicates the (approximate) minimum size of variants each method is able to detect, except when a range is given, where the maximum is also indicated. Note that not all methods are suitable for all CNVs; further details are given in the text

long read sequencing methods will also provide "gold standard" CNV calls for calibrating other calls.

Targeted Y-CNV studies

The *AZF* loci on the long arm are among the most active rearrangement hotspots in the human genome, and are some of the most studied because of their medical relevance (Repping et al. 2006). The three loci (*AZFa*, *AZFb* and *AZFc*, with *AZFb* and *AZFc* partially overlapping) were identified when their deletion was associated with azoospermia or severe oligozoospermia (Vogt et al. 1996). In recent years, *AZF* rearrangements have been surveyed in samples from many populations, including the Chinese (Lu et al. 2009, 2011; Yang et al. 2015; Zhang et al. 2013; Zhu et al. 2016), Dutch (Noordam et al. 2011), Europeans (Krausz et al. 2009), Jordanians (Khabour et al. 2014), Indians (Ambulkar et al. 2015; Kumari et al. 2015) and Iranians (Alimardanian et al. 2016; Motovali-Bashi et al. 2015), benefiting from the development of novel typing methods (Motovali-Bashi et al. 2015; Saito et al. 2015; Zhou et al. 2015; Zhu et al. 2016). The effects of *AZF* rearrangements on fertility have been summarised in a study by Lo Giacco et al. (2014), which presented data collected from diagnostic infertility testing over several years. Among 806 sterile males from several populations (~73% Spanish), the authors report 27 males with complete *AZF* deletion, including six showing abnormal karyotype and 21 with Y chromosome microdeletion. The authors also conducted a case–control study of partial *AZFc* deletions, showing that *AZFc* gr/gr and b2/b3 deletions (Repping et al. 2003), discussed further below, where significantly more frequent among sterile males than controls.

Among the *AZF* loci, *AZFc* region stands out for its complexity (Kuroda-Kawaguchi et al. 2001) and for the variety of alternative structures (Lu et al. 2011; Repping et al. 2006; Yang et al. 2015). A large study was conducted on over 20,000 males from India, Poland, Tunisia, the United States and Vietnam, assaying sequence-tagged sites (STSs) that mark different microdeletions (Rozen et al. 2012). This survey found that 3.7% of the sample had one of four deletions (gr/gr, b1/b3, b2/b3 and b2/b4) among those previously described in the region (Kuroda-Kawaguchi et al. 2001; Reijo et al. 1995; Repping et al. 2002, 2003). Individual frequencies for the assayed deletions varied widely across populations, from 15% for the gr/gr deletion in Vietnamese males, to 0.043% for the b1/b3 and b2/b4 deletions in Polish individuals, and down to the undetectability of the P5/P1 and P4/P1 (*AZFbc*) deletions. Moreover, the frequency of gr/gr and b2/b3 varied significantly across populations, with the latter probably due to differences in the prevalence of haplogroup N1 samples, in which the deletion is fixed (Fernandes et al. 2004). Rozen and colleagues also observed that haplogroup R1a appeared enriched in gr/gr deletions in the Polish population and in b1/b3 deletions among the samples from the United States. Finally, they estimated population frequency and contribution to severe spermatogenic failure (SSF) for the gr/gr, b1/b3, and b2/b4 deletions, concluding that about 8% of cases of SSF could be explained by either the 3.5 Mb b2/b4 deletion, which is rare (0.043% frequency) but has a strong effect (145-fold risk increase), or the more common 1.6 Mb gr/gr deletion (2.2% frequency), which doubles the risk of SSF (Rozen et al. 2012). From this and other studies we see how the gr/gr deletion appears to have a major impact on fertility, due to its combination of frequency and risk increase. This result also emerges from several meta-analyses (Navarro-Costa et al. 2010; Stouffs et al. 2011; Tuttelmann et al. 2007; Visser et al. 2009). The different penetrance and variable effect on the risk of spermatogenic failure observed for *AZFc* deletions might also depend on co-occurring compensatory duplications, hinting that besides causing an imbalance in gene dosage, *AZFc* rearrangements might affect fertility by altering the non-coding structure of the region (Yang et al. 2015).

Another highly active rearrangement hotspot lies on the *p* (short) arm of the chromosome, where the *TSPY* gene is present in a large and highly variable number of copies, organized as an array of 20.4 kb long elements (*TSPY* major), plus a single copy of the gene (*TSPY* minor) located more distally (Skaletsky et al. 2003). The copy number of TSPY has been observed to vary widely across population samples, up to 64 copies (Mathias et al. 1994; Oakey and Tyler-Smith 1990; Repping et al. 2006; Tyler-Smith et al. 1988); intraspecific variation comparable to that in humans has also been recently observed in gorillas (Tomaszkiewicz et al. 2016). *TSPY* organization represents 70% of the differences in functional gene number between the Y chromosome of humans and chimpanzees (Hughes et al. 2010): while the overall *TSPY* copy number, including inactive copies, is similar, chimpanzee Y chromosomes carry three arrays rather than two, and most of the copies are pseudogenes. A more detailed human–chimpanzee comparison (Xue and Tyler-Smith 2011) suggested that an ancestral array in the human–chimpanzee common ancestor might have undergone expansion in the human lineage and multiple duplications in the chimpanzee lineage; moreover, human–chimpanzee sequence comparison pointed to positive selection as a likely mechanisms of evolution for *TSPY* (Xue and Tyler-Smith 2011), implying a selective advantage in having multiple copies of the gene. In humans, *TSPY* is expressed exclusively in testis (Schnieders et al. 1996), and its copy number has been shown to have an effect on spermatogenesis, although results on this are

discordant (Giachini et al. 2009; Nickkholgh et al. 2010; Vodicka et al. 2007): perhaps too few and too many copies both increase the chance of spermatogenic failure. All of these medically important Y-CNVs are shown in Fig. 1.

TSPY arrays are also involved in a different form of structural variation. Non-allelic homologous recombination (NAHR) between *TSPY* major and *TSPY* minor can cause a deletion over 3 Mb long (Jobling et al. 2007; Santos et al. 1998), which removes several genes. Among these, *AMELY* is probably the most studied due to its importance in forensics. Multiplex PCR coamplification of portions of different length of the *AMELX* and *AMELY* gene pair (106 and 112 bp, respectively) is routinely used for sex identification (Sullivan et al. 1993); however, *AMELY* deletions (including but not limited to *TSPY*-mediated deletions) or point mutations at primer binding sites cause the test to identify such males as females (Tozzo et al. 2013). The wide usage of *AMELX/Y* testing has then led to the discovery of different rearrangements involving *AMELY*, with frequencies ranging from around 0.02% (Chen et al. 2014; Mitchell et al. 2006; Xie et al. 2014) to 8% (Santos et al. 1998), and ranging in length between 304 bp (Mitchell et al. 2006) and 4 Mb (Jobling et al. 2007). Systematic studies of these variants are scarce; such rearrangements appear rare, except in South Asia where they reach ~2% (Thangaraj et al. 2002). Despite the co-deletion of *PRKY*, *TBL1Y* and *PCDH11Y* in the larger events (Jobling et al. 2007), and reciprocal duplications (Murphy et al. 2007; Wei et al. 2015), no phenotypic effect has been described so far besides the aforementioned sex testing failure. Novel methods for sex testing have been proposed, which often integrate or replace the *AMELX/Y* test with *UTX/Y*, *SRY* or microsatellite typing (Cadamuro et al. 2015; Santos et al. 1998; Tozzo et al. 2013).

Chromosome-wide studies

The recent escalation in large-scale genomics has produced a wealth of information that can be employed in CNV studies. A good example of this, and also of how several studies on copy number variation have excluded the Y chromosome from their analyses, is the study published by Johansson et al. (2015). This study used SNP-array data to analyse a total of 1718 males from 13 previously published projects which excluded the Y chromosome from their analyses, although it included 510 males from phase 3 of the HapMap project (International HapMap C 2003). The full dataset covered several different populations, and given the multiple origins of the data, included samples gathered for the purpose of analysing conditions as diverse as schizophrenia, bipolar disorder, developmental disorders, high-altitude adaptation, cancer prostate, motor neuron disease, and colorectal cancer. Some highly variable regions on the chromosome were covered incompletely (*AZFc*) or not covered at all (*TSPY*) by the SNP-array probes. Nevertheless, Johansson and colleagues were able to identify 25 Y-chromosomal CNV patterns in their sample set, with an excess of duplications over deletions. Some of the variants identified were novel, and three variants were extremely rare, being identified in one individual each. The authors reported a significant association of ten variants with one or more haplogroups, which might represent a signature of rare events, likely to happen once in the Y phylogeny. The authors also tested the association of CNVs with the conditions present in their dataset, but did not detect any significant association.

Large projects such as the DDD project (The Deciphering Developmental Disorders Study 2015) and the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010, 2012, 2015) are powerful enough to alone enable researchers to investigate copy number variation across the whole chromosome, as demonstrated by several studies in recent years. In one study published in 2015, CNVs across the whole MSY were investigated in 411 apparently healthy males from the UK, using an array CGH design that had been employed in the DDD study; SNP-array data were used to validate the CNVs discovered in a subset of individuals (Wei et al. 2015). After merging overlapping CNVs called in individual samples into CNV events (CNVEs) and manual curation, 22 curated CNVEs (curCNVEs) were identified. Raw, individual events ranged in length from less than 1 kb to over 3 Mb, the latter corresponding to the *AMELY* duplication described above. More than half the events were observed in just one individual, but six had frequency higher than 5%, up to 26% (107/411 individuals). Deletions (relative to the reference used) were more abundant than duplications, but this was heavily influenced by two curCNVEs that were deleted in 76 and 68 individuals, respectively. None of the ten curated CNV events present in more than one individual was specific to a single Y haplogroup, implying recurrent mutational events for all of them. The curated set of variants covered 24 protein-coding genes, some of which had already been extensively investigated for CNVs, like the *AZFc* region, *TSPY* and *AMELY* CNVs discussed above. In addition, a previously undocumented partial duplication in the *AZFa* region that also extends to the *UTY* gene, and frequent variation in the *RBMY* and *PRY* multicopy gene families, was presented.

In the same year, a study of Y-CNVs inferred from sequence data in samples from the 1000 Genomes pilot phase (Espinosa et al. 2015) was published. The sample set consisted of 70 males from four populations (YRI, CEU, CHB and JPT) sequenced at 2.3× average depth; ten samples at variable depth, obtained by merging sequencing

data from subsets of the same males belonging to the same haplogroup; and eight samples from the Complete Genomics Public Data set (v36 v2.0.0), at high (25.4×) depth. CNVs were mainly identified using a custom sequencing depth analysis, where the threshold and window size to be used were fine-tuned by comparing the full data for the reference sample (NA12891 from the CEU populations) to subsets of data from the same sample, varying said parameters and assuming that no CNV should be discovered in this case. To account for uncertainty in breakpoint definition, variants were merged if separated by 5 kb or less. This approach was complemented by the analysis of paired-end data available for some of the samples, and SNP array data and PCR amplifications were used to validate the full set; two variants discovered, but not validated, in previous studies (Mills et al. 2011; The 1000 Genomes Project Consortium 2012) were also validated and included in the final set. In total, 19 CNVs were reported, with 12 of these (63%) overlapping segmental duplication: again, repetitive regions appeared to be involved in the majority of rearrangements. A bias was observed towards the detection of larger events, as well as towards deletions over duplications. The samples in this study belonged to ten different Y haplogroups; by leveraging the univocal phylogeny available for the Y chromosome, the minimum number of mutational events for each CNV was determined: out of the 19 variants, four appeared to be caused by single events, while 15 appeared to be due to multiple mutations. A possible explanation of this imbalance is the different contribution of mutational mechanisms involved in CNV formation, namely non-homologous end joining (NHEJ) and NAHR, with homology-mediated mechanisms being more prone to recurrent events than non-homology mechanisms. Alternative allele (i.e. non-reference allele) count varied between one and 64; most of the variants were located in ampliconic or heterochromatic regions (8/19 and 7/19, respectively), with the latter being also associated with most of the high alternative allele counts. Six CNVs overlapped with members of five gene families on the chromosome (*BPY*, *CDY*, *DAZ*, *PRY*, *TSPY*). Common variants known to be present in the analysed populations (Jobling et al. 1996; Oakey and Tyler-Smith 1990; Redon et al. 2006; Tyler-Smith et al. 1988) were all observed in this study, showing that NGS data, even at low depth, can be successfully used to investigate Y-CNVs.

In its third and final phase, the 1000 Genomes Project increased its samples size to 2504 individuals from 26 populations, and its mean whole-genome depth to 7.4× (The 1000 Genomes Project Consortium 2015). These resources enabled a large Y chromosome study to be carried out, which represents the widest description of MSY variation so far (Poznik et al. 2016). This work tackled all aspects of the MSY diversity: sequencing data for 1,244 males were used to discover over ~60,000 SNPs, which then were used to reconstruct an extensive phylogenetic tree; variants in other classes, including indels and multiple nucleotide polymorphisms (MNPs), CNVs and short tandem repeats (STRs), were discovered as well, and projected onto the high-resolution phylogeny to investigate their mutational patterns and properties. The main discovery method for CNVs was again the analysis of sequence data, using Genome STRiP (Handsaker et al. 2015), which infers structural rearrangements using the full information available from population-scale sequence data: local read depth variation, abnormal paired-end insert length, breakpoint-spanning reads, allele and haplotype sharing between samples, population heterogeneity caused by variant alleles, and negative correlation between alternative alleles (referred to as "allelic substitution") (Handsaker et al. 2011). This approach was complemented by array CGH data, which were used to validate the Genome STRiP set, and call additional variants, for a total of 121 CNVs reported (100 in the Genome STRiP set only). A set of variants was validated using alkaline lysis fibre-FISH and molecular combing fibre-FISH, together with karyotyping for samples showing sex chromosomes aneuploidies. The unbiased phylogeny reconstructed in the study was leveraged to count the minimum number of mutational events for each locus (Fig. 3): the majority of variants were explained as single mutational events, although a few loci showed evidence for a high number of mutations; there was a higher prevalence of duplications compared to deletions. The presence of repetitive elements near putative breakpoints did not appear to be associated with highly mutable loci, although it appeared to be associated with longer variants, similar to observations on the autosomes (Conrad et al. 2010) (Fig. 1). Unsurprisingly, CNVs were predicted to have larger phenotypic effect than single nucleotide variants, as inferred from overlap with protein-coding genes: however, deletions overlapping protein-coding genes appeared to be more abundant than duplications overlapping protein-coding genes, while in a reanalysis of 1000 Genomes Project autosomal data (Sudmant et al. 2015) this relation appeared to be reversed. In other words, Y genes appeared to be more tolerant to deletions than autosomal genes, despite the haploidy of the chromosome, probably owing to their presence in multiple copies in many cases.

Conclusions

The view of a Y chromosome confined to determining male fertility is being gradually superseded (Hughes and Page 2015), although the full understanding of copy number variation mechanisms and its impact on human biology is far from complete (Huddleston and Eichler 2016). Targeted

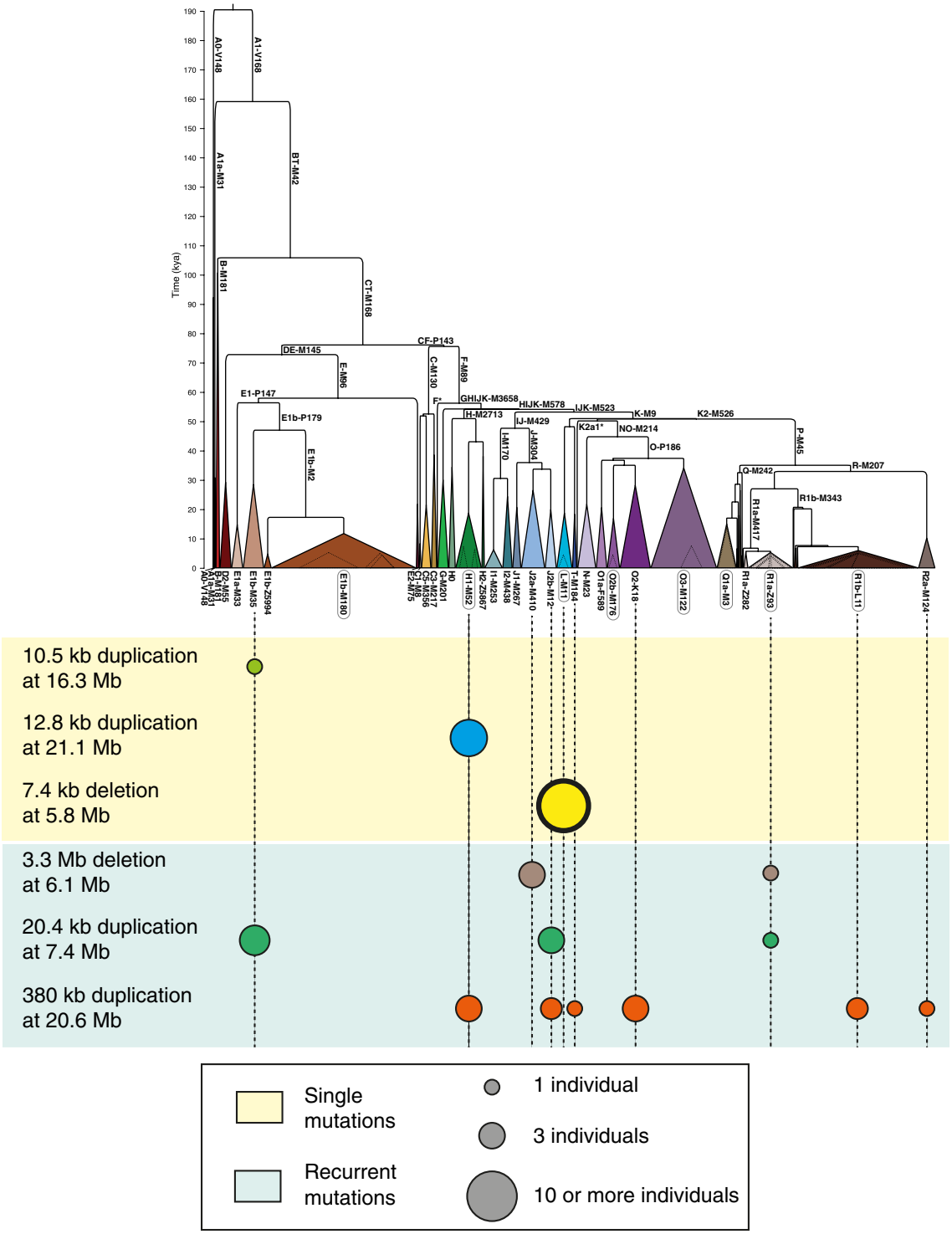


Fig. 3 Numbers of CNV mutational events inferred from the phylogenetic tree. **a** Phylogeny based on the 1000 Genomes Project phase 3 data (from Poznik et al. 2016). **b** Examples of single mutational event CNVs (*light yellow background*) and multiple event CNVs (*light green background*). The 7.4 deletion at 5.8 Mb CNV, with a thicker surround, indicates that all sampled members of this haplogroup carry this CNV, while in the other examples only some member(s) of the haplogroup carry the CNV. For each listed CNV, approximate chromosomal position in GRCh37 is given by, e.g. 'at 16.3 Mb'

studies, especially related to fertility, will likely continue to be carried out, perhaps on larger, population-specific cohorts, towards a complete description of variation in complex regions such as the *AZF* loci. Meanwhile, chromosome-wide studies should continue to uncover the full structural variation on the Y, filling the gaps and describing mutational mechanisms. It seems likely that almost all euchromatic Y-CNVs larger than 20 kb in the MSY that are frequent or fixed in any haplogroup have already been detected. Nevertheless, vast numbers of smaller and rarer Y-CNVs undoubtedly remain to be discovered, and study of the highly repeated highly variable heterochromatic segments has barely begun (Mathias et al. 1994). Future directions for Y-CNV investigations seem to lead towards the integration of different methods, especially with the developments of the long read technologies. A first attempt of such integration has been used to assemble the sequence of the gorilla Y chromosome (Tomaszkiewicz et al. 2016). This project employed flow-sorting (Dolezel et al. 2012) to obtain ~12,000 copies of the gorilla Y. These were used in a combination of short- and long-insert short read (Illumina Paired-End and Illumina Mate Pair sequencing, respectively) and long read sequencing (PacBio). Specific computational approaches were developed to increase the detection of Y-specific reads, which were used for a multi-step *de novo* assembly. RNA-seq from testis was employed to refine gene identification, and the size of ampliconic gene families was estimated using droplet digital PCR (Hindson et al. 2011). At least, until technologies leap forward once more, allowing cheap and accurate sequencing of Mb-sized molecules, complementing the weaknesses of one approach with the strengths of another is an attractive way to go.

Acknowledgements AM and YLX are supported by the Wellcome Trust (098051).

Compliance with ethical standards

Disclosure of potential conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Box 1. The shadowing effect

Methods based on mapping data to the reference genome are affected by a limitation called "shadowing". This phenomenon affects both intensity data, such as those produced by an array CGH experiment, and sequencing data, where the read depth interpretation is affected. Shadowing is caused by the presence of highly similar intra-chromosomal or inter-chromosomal duplicated sequences in the reference genome, and as such, is particularly relevant in repeat-rich regions. In such duplicated regions, experimental data map equally well to each of the reference copies. When a duplication or deletion of one of the copies occurs, the increase or decrease in signal will be averaged across all the reference copies. This means that the copy number variation will then be harder to identify, as it will result in a reduced signal difference compared to the copy number variation of a unique region; moreover, it will be impossible to tell which of the multiple copies in the reference genome has been duplicated or deleted. On the Y chromosome, shadowing is particularly relevant in ampliconic sequences, especially in palindromes, and in X–Y duplicated regions.

References

Alimardanian L, Saliminejad K, Razi S, Ahani A (2016) Analysis of partial azoospermia factor c deletion and DAZ copy number in azoospermia and severe oligozoospermia. Andrologia 48:890–894. doi:10.1111/and.12527

Ambulkar P, Chuadhary A, Waghmare J, Tarnekar A, Pal A (2015) Prevalence of Y Chromosome microdeletions in idiopathic azoospermia cases in Central Indian Men. J Clin Diagn Res 9:GC01-4. doi:10.7860/JCDR/2015/15249.6515

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, Kremitzki C, Fulton RS, Dugan S, Ding Y, Morton D, Khan Z, Lewis L, Buhay C, Wang Q, Watt J, Holder M, Lee S, Nazareth L, Alfoldi J, Rozen S, Muzny DM, Warren WC, Gibbs RA, Wilson RK, Page DC (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508:494–499. doi:10.1038/nature13206

Betran E, Demuth JP, Williford A (2012) Why chromosome palindromes? Int J Evol Biol 2012:207958. doi:10.1155/2012/207958

Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Newburger DE, West R, Sidow A, Batzoglou S (2015) Read clouds uncover variation in complex regions of the human genome. Genome Res 25:1570–1580. doi:10.1101/gr.191189.115

Boon EM, Schlecht HB, Martin P, Daniels G, Vossen RH, den Dunnen JT, Bakker B, Elles R (2007) Y chromosome detection by Real Time PCR and pyrophosphorolysis-activated polymerisation using free fetal DNA isolated from maternal plasma. Prenat Diagn 27:932–937. doi:10.1002/pd.1804

Cadamuro VC, Bouakaze C, Croze M, Schiavinato S, Tonasso L, Gerard P, Fausser JL, Gibert M, Dugoujon JM, Braga J, Balaresque P (2015) Determined about sex: sex-testing in 45 primate species using a 2Y/1X sex-typing assay. Forensic Sci Int Genet 14:96–107. doi:10.1016/j.fsigen.2014.09.010

Chen W, Wu W, Cheng J, Zhang Y, Chen Y, Sun H (2014) Detection of the deletion on Yp11.2 in a Chinese population. Forensic Sci Int Genet 8:73–79. doi:10.1016/j.fsigen.2013.07.003

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control C, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712. doi:10.1038/nature08516

Dolezel J, Vrana J, Safar J, Bartos J, Kubalakova M, Simkova H (2012) Chromosomes in the flow to simplify genome analysis. Funct Integr Genom 12:397–416. doi:10.1007/s10142-012-0293-0

Espinosa JR, Ayub Q, Chen Y, Xue Y, Tyler-Smith C (2015) Structural variation on the human Y chromosome from population-scale resequencing. Croat Med J 56:194–207

Fernandes S, Paracchini S, Meyer LH, Floridia G, Tyler-Smith C, Vogt PH (2004) A large AZFc deletion removes DAZ3/DAZ4 and nearby genes from men in Y haplogroup N. Am J Hum Genet 74:180–187. doi:10.1086/381132

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands C, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PI, Sunyaev SR (2015) Genome-wide patterns and properties of de novo mutations in humans. Nat Genet 47:822–826. doi:10.1038/ng.3292

Giachini C, Nuti F, Turner DJ, Laface I, Xue Y, Daguin F, Forti G, Tyler-Smith C, Krausz C (2009) TSPY1 copy number variation influences spermatogenesis and shows differences among Y lineages. J Clin Endocrinol Metab 94:4016–4022. doi:10.1210/jc.2009-1029

Hammer MF (1994) A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol 11:749–761

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269–276. doi:10.1038/ng.768

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA (2015) Large multiallelic copy number variations in humans. Nat Genet. doi:10.1038/ng.3200

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, Renkens I, Coe BP, Deelen P, de Ligt J, Lameijer EW, van Dijk F, Hormozdiari F, Genome of the Netherlands C, Uitterlinden AG, van Duijn CM, Eichler EE, de Bakker PI, Swertz MA, Wijmenga C, van Ommen GB, Slagboom PE, Boomsma DI, Schonhuth A, Ye K, Guryev V (2016) A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat Commun 7:12989. doi:10.1038/ncomms12989

Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, Kitano TK, Hodel MR, Petersen JF, Wyatt PW, Steenblock ER, Shah PH, Bousse LJ, Troup CB, Mellen JC, Wittmann DK, Erndt NG, Cauley TH, Koehler RT, So AP, Dube S, Rose KA, Montesclaros L, Wang S, Stumbo DP, Hodges SP, Romine S, Milanovich FP, White HE, Regan JF, Karlin-Neumann GA, Hindson CM, Saxonov S, Colston BW (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83:8604–8610. doi:10.1021/ac202028g

Huddleston J, Eichler EE (2016) An incomplete understanding of human genetic variation. Genetics 202:1251–1254. doi:10.1534/genetics.115.180539

Hughes JF, Page DC (2015) The biology and evolution of mammalian Y chromosomes. Annu Rev Genet 49:507–527. doi:10.1146/annurev-genet-112414-055311

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, Trask BJ, Mardis ER, Warren WC, Repping S, Rozen S, Wilson RK, Page DC (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463:536–539. doi:10.1038/nature08700

International HapMap C (2003) The International HapMap Project. Nature 426:789–796. doi:10.1038/nature02168

Jacobs PA, Brunton M, Court Brown WM, Doll R, Goldstein H (1963) Change of human chromosome count distribution with age: evidence for a sex differences. Nature 197:1080–1081

Jobling MA (2008) Copy number variation on the human Y chromosome. Cytogenet Genome Res 123:253–262. doi:10.1159/000184715

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, Dashnyam B, Sajantila A, Salo PJ, Nakahori Y, Disteche CM, Thangaraj K, Singh L, Crawford MH, Tyler-Smith C (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767–1775

Jobling MA, Lo IC, Turner DJ, Bowden GR, Lee AC, Xue Y, Carvalho-Silva D, Hurles ME, Adams SM, Chang YM, Kraaijenbrink T, Henke J, Guanti G, McKeown B, van Oorschot RA, Mitchell RJ, de Knijff P, Tyler-Smith C, Parkin EJ (2007) Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y. Hum Mol Genet 16:307–316. doi:10.1093/hmg/ddl465

Johansson MM, Van Geystelen A, Larmuseau MH, Djurovic S, Andreassen OA, Agartz I, Jazin E (2015) Microarray analysis of copy number variants on the human Y chromosome reveals novel and frequent duplications overrepresented in specific haplogroups. PLoS One 10:e0137223. doi:10.1371/journal.pone.0137223

Khabour OF, Fararjeh AS, Alfaouri AA (2014) Genetic screening for AZF Y chromosome microdeletions in Jordanian azoospermic infertile men. Int J Mol Epidemiol Genet 5:47–50

Krausz C, Giachini C, Xue Y, O'Bryan MK, Gromoll J, Rajpert-de Meyts E, Oliva R, Aknin-Seifer I, Erdei E, Jorgensen N, Simoni M, Ballesca JL, Levy R, Balercia G, Piomboni P, Nieschlag E, Forti G, McLachlan R, Tyler-Smith C (2009) Phenotypic variation within European carriers of the Y-chromosomal gr/gr deletion is independent of Y-chromosomal background. J Med Genet 46:21–31. doi:10.1136/jmg.2008.059915

Kumari A, Yadav SK, Ali S (2012) Organizational and functional status of the Y-linked genes and loci in the infertile patients having normal spermiogram. PLoS One 7:e41488. doi:10.1371/journal.pone.0041488

Kumari A, Yadav SK, Misro MM, Ahmad J, Ali S (2015) Copy number variation and microdeletions of the Y chromosome linked genes and loci across different categories of Indian infertile males. Sci Rep 5:17780. doi:10.1038/srep17780

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, Page DC (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 29:279–286. doi:10.1038/ng757

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 3:1–8. doi:10.1016/j.bdq.2015.02.001

Legro RS (2012) Turner syndrome: new insights into an old disorder. Fertil Steril 98:773–774. doi:10.1016/j.fertnstert.2012.07.1138

Lo Giacco D, Chianese C, Sanchez-Curbelo J, Bassas L, Ruiz P, Rajmil O, Sarquella J, Vives A, Ruiz-Castane E, Oliva R, Ars E, Krausz C (2014) Clinical relevance of Y-linked CNV screening

in male infertility: new insights based on the 8-year experience of a diagnostic genetic laboratory. Eur J Hum Genet 22:754–761. doi:10.1038/ejhg.2013.253

Lu C, Zhang J, Li Y, Xia Y, Zhang F, Wu B, Wu W, Ji G, Gu A, Wang S, Jin L, Wang X (2009) The b2/b3 subdeletion shows higher risk of spermatogenic failure and higher frequency of complete AZFc deletion than the gr/gr subdeletion in a Chinese population. Hum Mol Genet 18:1122–1130. doi:10.1093/hmg/ddn427

Lu C, Zhang F, Yang H, Xu M, Du G, Wu W, An Y, Qin Y, Ji G, Han X, Gu A, Xia Y, Song L, Wang S, Jin L, Wang X (2011) Additional genomic duplications in AZFc underlie the b2/b3 deletion-associated risk of spermatogenic impairment in Han Chinese population. Hum Mol Genet 20:4411–4421. doi:10.1093/hmg/ddr369

Mathias N, Bayes M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO, Genomes P (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65. doi:10.1038/nature09708

Mitchell RJ, Kreskas M, Baxter E, Buffalino L, Van Oorschot RA (2006) An investigation of sequence deletions of amelogenin (AMELY), a Y-chromosome locus commonly used for gender determination. Ann Hum Biol 33:227–240. doi:10.1080/03014460600594620

Motovali-Bashi M, Rezaei Z, Dehghanian F, Rezaei H (2015) Multiplex PCR based screening for micro/partial deletions in the AZF region of Y-chromosome in severe oligozoospermic and azoospermic infertile men in Iran. Iran J Reprod Med 13:563–570

Murphy KM, Cohen JS, Goodrich A, Long PP, Griffin CA (2007) Constitutional duplication of a region of chromosome Yp encoding AMELY, PRKY, and TBL1Y: implications for sex chromosome analysis and bone marrow engraftment analysis. J Mol Diagn 9:408–413. doi:10.2353/jmoldx.2007.060198

Navarro-Costa P, Goncalves J, Plancha CE (2010) The AZFc region of the Y chromosome: at the crossroads between genetic diversity and male infertility. Hum Reprod Update 16:525–542. doi:10.1093/humupd/dmq005

Nickkholgh B, Noordam MJ, Hovingh SE, van Pelt AM, van der Veen F, Repping S (2010) Y chromosome TSPY copy numbers and semen quality. Fertil Steril 94:1744–1747. doi:10.1016/j.fertnstert.2009.09.051

Noordam MJ, Westerveld GH, Hovingh SE, van Daalen SK, Korver CM, van der Veen F, van Pelt AM, Repping S (2011) Gene copy number reduction in the azoospermia factor c (AZFc) region and its effect on total motile sperm count. Hum Mol Genet 20:2457–2463. doi:10.1093/hmg/ddr119

Oakey R, Tyler-Smith C (1990) Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. Genomics 7:325–330

Paoloni-Giacobino A, Lespinasse J (2007) Chromosome Y polysomy: a non-mosaic 49, XYYYY case. Clin Dysmorphol 16:65–66. doi:10.1097/01.mcd.0000228423.04908.0c

Pirooznia M, Goes FS, Zandi PP (2015) Whole-genome CNV analysis: advances in computational approaches. Front Genet 6:138. doi:10.3389/fgene.2015.00138

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X, Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE, Genomes Project C, Bustamante CD, Tyler-Smith C (2016) Punctuated bursts in human male demography inferred from 1244 worldwide Y-chromosome sequences. Nat Genet 48:593–599. doi:10.1038/ng.3559

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. Nature 444:444–454. doi:10.1038/nature05329

Reijo R, Lee TY, Salo P, Alagappan R, Brown LG, Rosenberg M, Rozen S, Jaffe T, Straus D, Hovatta O et al (1995) Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. Nat Genet 10:383–393. doi:10.1038/ng0895-383

Repping S, Skaletsky H, Lange J, Silber S, Van Der Veen F, Oates RD, Page DC, Rozen S (2002) Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. Am J Hum Genet 71:906–922. doi:10.1086/342928

Repping S, Skaletsky H, Brown L, van Daalen SK, Korver CM, Pyntikova T, Kuroda-Kawaguchi T, de Vries JW, Oates RD, Silber S, van der Veen F, Page DC, Rozen S (2003) Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nat Genet 35:247–251. doi:10.1038/ng1250

Repping S, van Daalen SK, Brown LG, Korver CM, Lange J, Marszalek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, Rozen S (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet 38:463–467. doi:10.1038/ng1754

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genom Proteom Bioinform 13:278–289. doi:10.1016/j.gpb.2015.08.002

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876. doi:10.1038/nature01723

Rozen SG, Marszalek JD, Irenze K, Skaletsky H, Brown LG, Oates RD, Silber SJ, Ardlie K, Page DC (2012) AZFc deletions and spermatogenic failure: a population-based survey of 20,000 Y chromosomes. Am J Hum Genet 91:890–896. doi:10.1016/j.ajhg.2012.09.003

Saito K, Miyado M, Kobori Y, Tanaka Y, Ishikawa H, Yoshida A, Katsumi M, Saito H, Kubota T, Okada H, Ogata T, Fukami M (2015) Copy-number variations in Y-chromosomal azoospermia factor regions identified by multiplex ligation-dependent probe amplification. J Hum Genet 60:127–131. doi:10.1038/jhg.2014.115

Santos FR, Pandya A, Tyler-Smith C (1998) Reliability of DNA-based sex tests. Nat Genet 18:103. doi:10.1038/ng0298-103

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A, Singh L, Destro-Bisol G, Novelletto A, Qamar R, Mehdi SQ, Adhikari R, de Knijff P, Tyler-Smith C (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9:421–430

Schnieders F, Dork T, Arnemann J, Vogel T, Werner M, Schmidtke J (1996) Testis-specific protein, Y-encoded (TSPY) expression in testicular tissues. Hum Mol Genet 5:1801–1807

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–837. doi:10.1038/nature01722

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, West RB, Batzoglou S, Sidow A (2016) Genome-wide reconstruction of complex structural variants using read clouds. bioRxiv. doi:10.1101/074518

Stouffs K, Lissens W, Tournaye H, Haentjens P (2011) What about gr/gr deletions and male infertility? Systematic review and meta-analysis. Hum Reprod Update 17:197–209. doi:10.1093/humupd/dmq046

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stutz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJ, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HY, Jasmine MuX, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2504 human genomes. Nature 526:75–81. doi:10.1038/nature15394

Sullivan KM, Mannucci A, Kimpton CP, Gill P (1993) A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. Biotechniques 15(636–8):640–641

Thangaraj K, Reddy AG, Singh L (2002) Is the amelogenin gene reliable for gender identification in forensic casework and prenatal diagnosis? Int J Legal Med 116:121–123

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. doi:10.1038/nature09534

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1092 human genomes. Nature 491:56–65. doi:10.1038/nature11632

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74. doi:10.1038/nature15393

The Deciphering Developmental Disorders Study (2015) Large-scale discovery of novel genetic causes of developmental disorders. Nature 519:223–228. doi:10.1038/nature14135

The Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46:818–825. doi:10.1038/ng.3021

Tiepolo L, Zuffardi O (1976) Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human Y chromosome long arm. Hum Genet 34:119–124

Tomaszkiewicz M, Rangavittal S, Cechova M, Campos Sanchez R, Fescemyer HW, Harris R, Ye D, O'Brien PCM, Chikhi R, Ryder OA, Ferguson-Smith MA, Medvedev P, Makova KD (2016) A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. Genome Res. doi:10.1101/gr.199448.115

Tozzo P, Giuliodori A, Corato S, Ponzano E, Rodriguez D, Caenazzo L (2013) Deletion of amelogenin Y-locus in forensics: literature revision and description of a novel method for sex confirmation. J Forensic Leg Med 20:387–391. doi:10.1016/j.jflm.2013.03.012

Tuttelmann F, Rajpert-De Meyts E, Nieschlag E, Simoni M (2007) Gene polymorphisms and male infertility–a meta-analysis and literature review. Reprod Biomed Online 15:643–658

Tyler-Smith C, Taylor L, Muller U (1988) Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome. J Mol Biol 203:837–848

Visser L, Westerveld GH, Korver CM, van Daalen SK, Hovingh SE, Rozen S, van der Veen F, Repping S (2009) Y chromosome gr/gr deletions are a risk factor for low semen quality. Hum Reprod 24:2667–2673. doi:10.1093/humrep/dep243

Vodicka R, Vrtel R, Dusek L, Singh AR, Krizova K, Svacinova V, Horinova V, Dostal J, Oborna I, Brezinova J, Sobek A, Santavy J (2007) TSPY gene copy number as a potential new risk factor for male infertility. Reprod Biomed Online 14:579–587

Vogt PH, Bender U (2013) Human Y chromosome microdeletion analysis by PCR multiplex protocols identifying only clinically relevant AZF microdeletions. Methods Mol Biol 927:187–204. doi:10.1007/978-1-62703-038-0_17

Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Kohn FM, Schill WB, Farah S, Ramos C, Hartmann M, Hartschuh W, Meschede D, Behre HM, Castel A, Nieschlag E, Weidner W, Grone HJ, Jung A, Engel W, Haidl G (1996) Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum Mol Genet 5:933–943

Wei W, Fitzgerald T, Ayub Q, Massaia A, Smith BB, Dominiczak AA, Morris AA, Porteous DD, Hurles ME, Tyler-Smith C, Xue Y (2015) Copy number variation in the human Y chromosome in the UK population. Hum Genet 134:789–800. doi:10.1007/s00439-015-1562-5

Xie J, Shao C, Xu H, Zhu W, Liu Z, Tang Q, Zhou Y (2014) Deletion mapping of the regions with AMELY from two Chinese males. Leg Med (Tokyo) 16:290–292. doi:10.1016/j.legalmed.2014.05.002

Xue Y, Tyler-Smith C (2011) An Exceptional Gene: evolution of the TSPY Gene family in humans and other great apes. Genes (Basel) 2:36–47. doi:10.3390/genes2010036

Yang B, Ma YY, Liu YQ, Li L, Yang D, Tu WL, Shen Y, Dong Q, Yang Y (2015) Common AZFc structure may possess the optimal spermatogenesis efficiency relative to the rearranged structures mediated by non-allele homologous recombination. Sci Rep 5:10551. doi:10.1038/srep10551

Zhang YS, Dai RL, Wang RX, Zhang HG, Chen S, Liu RZ (2013) Analysis of Y chromosome microdeletion in 1738 infertile men from northeastern China. Urology 82:584–588. doi:10.1016/j.urology.2013.04.017

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM, Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP, Wu I, Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, Ji HP (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. doi:10.1038/nbt.3432

Zhou Y, Ge Y, Xiao L, Guo Q (2015) Rapid and simultaneous screening of 47, XXY and AZF microdeletions by quadruplex real-time polymerase chain reaction. Reprod Biol 15:113–121. doi:10.1016/j.repbio.2015.02.002

Zhu XB, Gong YH, He J, Guo AL, Zhi EL, Yao JE, Zhu BS, Zhang AJ, Li Z (2016) Multicentre study of Y chromosome microdeletions in 1808 Chinese infertile males using multiplex and real-time polymerase chain reaction. Andrologia. doi:10.1111/and.12662