Genetic and Epigenetic Basis of Development and Disease

3

Peter A. Kaub and Christopher P. Barnett

Abstract

This chapter aims to give an overview of the basis, forms, and output of genetic testing. It is intended to be a quick introductory reference and primer to more detailed sources, such as the references (predominantly reviews) and many online sources cited. Divided into two parts, the first section aims to outline genetic structures and their modes of inheritance to explain the genetic basis of disease. The second section gives an overview of the main technologies currently available for genetic testing, outlining the basic concepts underpinning each test, simple laboratory considerations, plus some commentary on result interpretation and limitations. Useful if read in its entirety, this chapter is also designed to allow easy reference and jumping between sections if only after a definition, refresher of theory, or specific details of a test, technology, or online database.

Keywords

Human genome • Genetic structures • Deoxyribonucleic acid (DNA) • Ribonucleic acid (RNA) • Chromosome • Ploidy • Cytogenetics • Karyotype • Mutation • Nomenclature Databases • Mendelian inheritance • Next-generation sequencing • Sex-linked inheritance • Epigenetics • Epigenome

The publication of the first draft map of the entire human genome in 2001 launched a new era in the field of genetic testing. Then, as now, the abundance of new data about an individual's genetic makeup has likely asked just as many questions as it has answered. Genetic testing laboratories still rely today on several more traditional, older techniques. However, we are likely at a crossroads, where the power and complexity of technologies already available are set to revolutionize diagnosis and management, with a profound impact on the practices of pathology and medicine. Soon it will be cheaper to sequence the entire genetic makeup of an individual than a single gene using older techniques.

Historical context is useful to fully appreciate the current pace of change in the sphere of genetic testing. Figure 3.1 shows a timeline of some of the major scientific and technological breakthroughs in this field, highlighting the very rapid pace of the evolving field of next-generation sequencing.

This chapter aims to give an overview of the basis, forms, and output of genetic testing. It is intended to be a quick introductory reference and primer to more detailed sources, such as the references (predominantly reviews) and many online sources cited. Divided into two parts, the first section aims to outline genetic structures and their modes of inheritance to explain the genetic basis of disease. The second section gives an overview of the main technologies currently available for genetic testing, outlining the basic concepts underpinning each test, simple laboratory considerations, and some commentary on result interpretation and

P.A. Kaub, BSc (Biotechnology) (Hons), MBBS (✉)
Genetics and Molecular Pathology,
SA Pathology, Women's and Children's Hospital,
Royal Adelaide Hospital and University of Adelaide,
Adelaide, SA, Australia
e-mail: Peter.Kaub@adelaide.edu.au; Peter.Kaub@health.sa.gov.au

C.P. Barnett, MBBS, FRACP, FCCMG
Paediatric and Reproductive Genetics Unit,
Women's and Children's Hospital,
North Adelaide, SA, Australia

T.Y. Khong, R.D.G. Malcomson (eds.), *Keeling's Fetal and Neonatal Pathology*, DOI 10.1007/978-3-319-19207-9_3

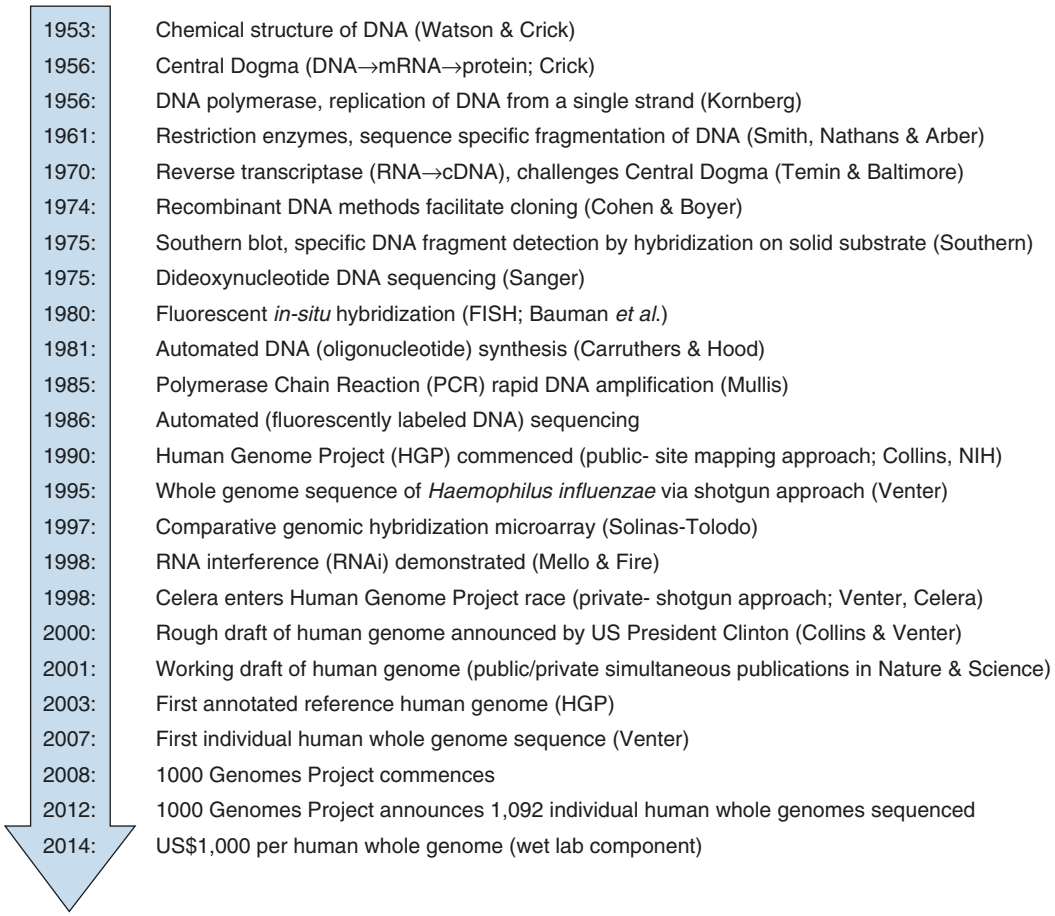


Fig. 3.1 Timeline of major scientific and technological milestones in genetic testing

limitations. Useful if read in its entirety, this chapter is also designed to allow easy reference and jumping between sections if only after a definition, refresher of theory, or specific details of a test, technology, or online database.

Clear definitions of nomenclature are necessary to navigate this complex and ever-expanding field. Keywords appear in bold type to enable easy identification where they are discussed or defined.

New genetic technologies are quickly emerging, with both output quality and cost continuing to improve rapidly. Therefore, genetic testing is likely to further pervade an increasing number of medical specialities (especially pathology). Just as scientific discovery is a process of "standing on the shoulders of giants," new genetic technologies build on previous scientific and technological breakthroughs. Therefore, a grasp of the underlying principles presented here will likely be useful for understanding new and emerging genetic technologies into the future.

Genetic Structures

To understand the mechanisms underlying genetic anomalies, an understanding of genetic structures is essential. Human cells contain a nucleus consisting of highly condensed nucleic acids, mostly deoxyribonucleic acid (DNA) with some ribonucleic acid (RNA), plus protein to form a unit called **chromatin**. Chromatin structure changes during the cell cycle to allow DNA replication and repair, as well as normal gene regulation and expression. Chromatin contains pairs of DNA containing **chromosomes** connected by a **centromere**, tightly bound around disks of **histone** (an alkaline protein), to form a **nucleosome**. In humans, chromosomes are classified into 22 pairs of **autosomes** (numbered chromosomes) and one pair of **allosomes** (sex chromosomes; XX female, XY male). **Chromosome number** is based on approximate size, with chromosome 1 being much larger than chromosome 22. **Ploidy** refers to the chromosome state; e.g., **diploid** for pairs of chromosomes, **haploid** for single chromosomes, and **aneuploid** for an incorrect multiple of chromosomes (e.g., triploid $n=3$, tetraploid $n=4$).

Produced through the process of **meiosis in the gonads**, gametes retain only one member of each pair of chromosomes (haploid; $n=1$). When gametes fuse in the process of conception to form a zygote, a paired complement (diploid; $n=2$) of chromosomes is formed.

Mitosis is the process of production of two daughter cells from a single cell. This is important for replication of cells in both growth and development, as well as maintenance of normal cell turnover throughout life. Each daughter cell

Fig. 3.2 DNA double helix. Attached to a sugar/phosphate backbone (*gray*), complementary nucleotides A and T (*green* and *red*) or G and C (*violet* and *blue*) bind to each other, like rungs on a rope ladder, in the tightly wound double-stranded helical structure of DNA. The specificity of this complementary binding gives DNA its information coding and high-fidelity replication abilities plus underpins the fundamental basis for the vast majority of DNA test technologies used today

contains identical copies of the full complement of chromosome pairs, tightly packed into a nucleus.

Cytogenetics classifies chromosomes according to well-characterized banding patterns, following special staining, to produce a **karyotype** (see page 59).

Recombination is a process whereby DNA is swapped across chromosomes. It happens during meiosis across **homologous chromosomes** (containing the same alleles) to produce new variations of haploid chromosomes in the gametes—a normal function of sexual reproduction that generates diversity in offspring. Recombination can also occur during mitosis as part of normal mechanisms of **homologous recombinational repair**, usually after damage is sustained to one allele. ***Non-homologous* recombination** can lead to disease from insertion of genes into inappropriate regions (e.g., translocation or inversion), a frequent mechanism underlying cancers.

Although there are many inbuilt checking and repair mechanisms, each of the aforementioned processes has the potential for introduction of changes into DNA. Generally, these are called **mutations** if detrimental or **variations** if not known to be detrimental. Humans share about 99.5 % of their DNA, with variations in the remaining small percentage responsible for the differences in specific traits or disease between individuals. Recent trends in nomenclature have fallen on the side of naming all DNA changes in an individual as **variants** (see SNVs on page 50).

Mutations may be inherited from parents (**germline**), generated during meiosis in sperm or ova (*de novo* mutations or **gonadal mosaicism** if more than one sperm or ovum carries the mutation), newly produced during the process of development of an embryo (***de novo***), or accumulated (**somatic**) from environmental exposure to chemicals, radiation, or toxins, or from normal accumulation of errors during the many cycles of replication and repair throughout life.

When unraveled, the chromosomes are found to consist mostly of a double-stranded helical structure of DNA (Fig. 3.2).

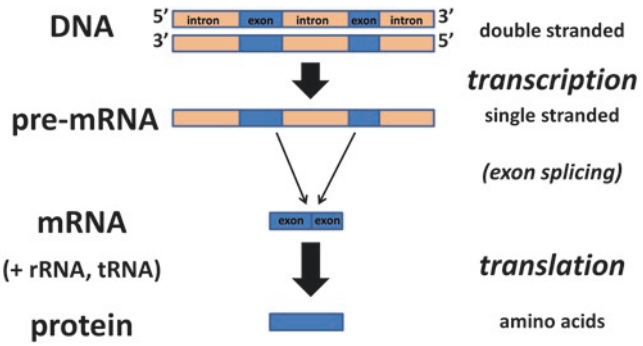


Fig. 3.3 Central dogma of genetics. In a one-way, linear fashion, information coded in double-stranded DNA is transcribed into messenger RNA (*mRNA*), which is then translated into protein (with the assistance of ribosomal and transfer RNAs: *rRNA* and *tRNA*, respectively). Although not part of Crick's original central dogma description, it was subsequently determined that a precursor *mRNA* (*pre-mRNA*) step is where introns are removed and exons spliced together to form the mature *mRNA* transcript

The chemical structure of DNA consists of a five-carbon (pentose) sugar (deoxyribose) with base organic **nucleotides** (**cytosine**, **adenine**, **thymine**, **guanine**; abbreviated **CATG**). There are two strands of DNA linked by phosphate groups in a double helical structure via the pairing of complementary nucleotides. C will only bind with G, and A will only bind with T (e.g., CGTACG will only bind with GCATGC).

Following on from his Nobel Prize-winning description of the chemical structure of DNA with James Watson, Francis Crick proposed the concept of the "**central dogma**" to explain how DNA impacts on cell and organism-level functioning. In this model, there is a one-way production of proteins (**translation**) via the intermediary of **messenger RNA** (**mRNA**) from the DNA blueprint (**transcription**) (Fig. 3.3). From this model also came the concept of the **gene**, i.e., a sequence of DNA responsible for producing a protein. While useful as a

simple explanation for the role of DNA, this model has subsequently been found to have many caveats, due to many other modifications now known to occur (e.g., reverse transcriptase enzyme in retroviruses allowing RNA to produce cDNA, plus the field of epigenetics, discussed on page 57). Currently a **gene** is defined by the HGNC (HUGO Gene Nomenclature Committee) as "a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology" [1]. This takes into account the concept of a gene, where it may play a role in modifying physiological function or regulation without explicit protein production or even without an immediately apparent functional process.

The entire sequence of nucleotides (**genome**) in humans consists of approximately 3.2 billion complementary nucleotide pairs (often called **base pairs** "**bp**") bound together in double helical strands. The two strands contain an antiparallel mirror of the sequence of each other, each nucleotide bound to its complementary pair on the opposing strand (Fig. 3.2).

Replication of DNA requires a tightly orchestrated process involving several enzymes. The ends of a DNA strand are denoted as 5′ (five prime) or 3′ (three prime), and DNA replication always proceeds in a 5–3′ direction. The enzyme topoisomerase acts to uncoil the densely packed DNA strands. DNA helicase, a motor protein, then breaks hydrogen bonds between the DNA strands separating them into single strands producing a replication fork, composed of a "leading" and a "lagging" strand (Fig. 3.4). The enzyme primase synthesizes a short RNA fragment (primer) that binds to the start of a region requiring replication. DNA polymerase is then able to add additional nucleotides to the 3′ end of the growing replicating strand. The leading strand is in the right orientation for continuous replication 5–3′; however, the lagging strand is in the reverse orientation and therefore requires the use of smaller "Okazaki fragments" to replicate, joined together by DNA ligase to make a continuous replicated strand. Very good animation of this process is abundant in free online video-sharing sites (the Cold Spring Harbor Laboratory DNA Learning Center is a good starting point [2]).

The process of replication is performed with very high fidelity, but errors still occur at a rate of approximately one in every 100,000 bp. In a genome of 3 billion bp, this can equate to up to 300,000 errors every time a cell divides. DNA polymerase itself has a proofreading mechanism that fixes about 99 % of these errors. **Mismatch repair** is a mechanism that monitors for kinks in DNA secondary structure caused by incorrectly incorporated non-complementary nucleotides, replacing them with a complementary nucleotide. While these processes are very robust, they can also cause introduction of errors in DNA sequence (**mutations**), which become permanent for all subsequent daughter cells.

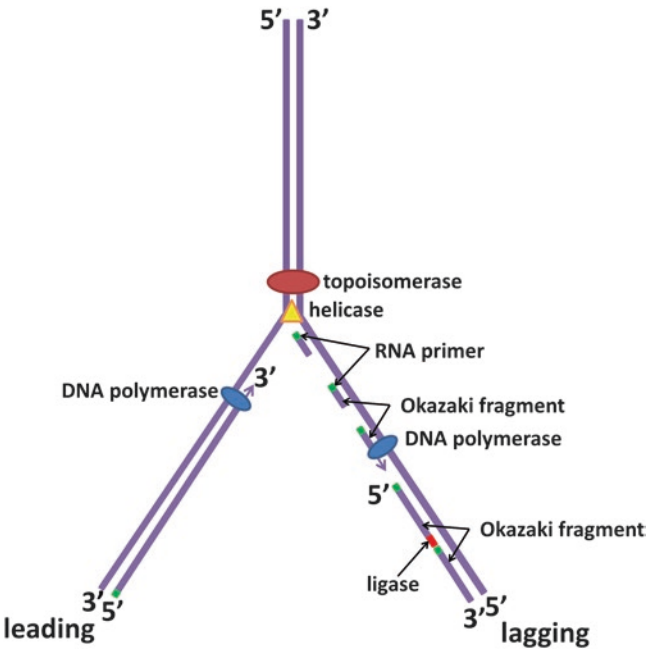


Fig. 3.4 DNA replication fork. By breaking the hydrogen bonds between complementary nucleotides and unwinding the DNA, enzymes topoisomerase and helicase combine to temporarily separate double-stranded DNA into single strands. This allows RNA primers produced by primase to anneal to complementary regions on target DNA. DNA polymerase binds next to these primers and makes a complementary copy of the single strand of DNA it is bound to, producing two molecules of double-stranded DNA. Replication can only occur in the 5–3′ prime direction, a simple process on the "leading strand." However, the "lagging strand" requires a different approach for 5–3′ replication, involving multiple RNA primers and piecewise production of "Okazaki fragments." The gaps between Okazaki fragments are then filled in by the enzyme DNA ligase

A three-nucleotide sequence (**codon**) and its relative alignment determine which amino acid will be **translated** into a growing protein chain, e.g., CAG for glutamine (a full codon usage table is available from the Human Genome Variation Society [3]). Mutations are classified according to the impact a nucleotide change has on translation to an amino acid. Translation to the same amino acid is called a **synonymous mutation**; translation to a different amino acid is called a **missense mutation** (non-synonymous); and if translation is stopped by the introduction of a stop codon, it is called a **nonsense mutation** (Table 3.1).

Mutations frequently occur for single nucleotides but can involve **in**sertions or **del**etions (concatenated to **indel**) of varying lengths. A change that occurs in greater than 1 % of the population is, by virtue of its prevalence, likely to be a normal variant and not pathogenic. **SNPs** (single-nucleotide polymorphisms) and **SNVs** (single-nucleotide variants) refer to single-nucleotide changes that occur at a population and an individual level, respectively. It is an SNV that may be unique to an individual and worth investigating for its role in disease. SNVs are classified as pathogenic, benign, or of unknown significance

Table 3.1 Classification of mutation types

Mutation type	DNA sequence	Amino acid (protein) change
Synonymous	CA**G** ➪ CA**A**	Glutamine ➪ glutamine
Missense	CC**C** ➪ CA**C**	Proline ➪ histidine
Nonsense	**A**AA ➪ **T**AA	Lysine ➪ stop codon (denoted by an asterisk * in sequence text)
Frameshift (+1 bp)	TGT-CAC ➪ TGT**G**CAC	Cysteine, histidine ➪ cysteine, alanine and following amino acids likely changed
Deletion (in-frame)	CAG**TGT**CAC ➪ CAG-CAC	Loss of cysteine (often important site for disulfide bonds) between glutamine and histidine

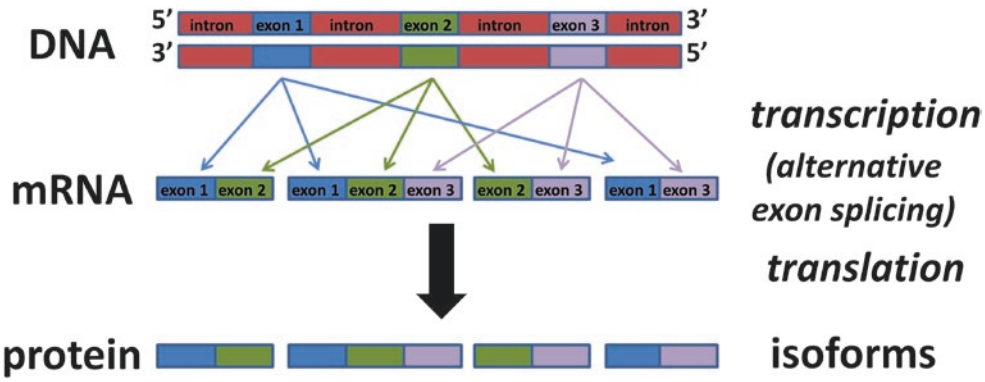


Fig. 3.5 Alternative splicing of exons. Differential splicing of intron/exon junctions can produce different combinations of exons in the mature mRNA transcript. This results in different isoforms of protein from the same gene. Failure to remove introns or incorrect splicing of number and/or order of exons can also lead to disease

(see Bioinformatics, page 81). A **frameshift mutation** involves insertion or deletion of one or more nucleotides that shift the reading frame of the following nucleotides so that the triplet codons now code for different amino acids. An **in-frame mutation** is when the number of nucleotides changed is an exact multiple of three. The amino acids before and after the mutation remain the same, but if the amino acid change is at an important structural position for protein folding or subcellular localization, then it is more likely to be detrimental. In-frame expansions are also important mechanisms in triplet repeat diseases, such as fragile X syndrome.

Approximately 99 % of the genome consists of regions that do not code for proteins. Much of this was previously thought to be "junk DNA," but evidence continues to emerge of regulatory and other roles of untranslated regions related to tissue-specific expression, e.g., non-coding RNA (see epigenetics on page 57). A protein-coding gene "edits" a large amount of information out in the process of transcription from DNA to mRNA. **Introns** are spliced out of the pre-mRNA and **exons** only are included in the mRNA transcript used for translation into protein (Fig. 3.3). **Alternative splicing** refers to a process whereby the incorporation or exclusion of different exons results in alternative sizes (**isoforms**) of a protein produced from the same gene (Fig. 3.5). The mechanisms involved in this process are too complex to detail in this brief chapter, but suffice to say, they are another potential source of the introduction of mutations (for review, see [4]).

Genome refers to the entire genetic complement of a species or individual. The entire complement of exons is referred to as the **exome,** and the entire complement of mRNA transcripts is referred to as the **transcriptome**. Similarly, the entire complement of proteins produced is called the **proteome**.

Mitochondrial DNA (**mtDNA**) is a separate entity to DNA in the nucleus (**nuclear DNA**). It is a circular, small, double-stranded entity of only 16.6 kbp, coding 37 known genes, associated with oxidative phosphorylation and translation regulation. Immensely important for energy (ATP) production, the mitochondria contain more than 1,500 proteins, with most coded by nuclear DNA and subsequently transported into the mitochondria. Importantly, the relatively small size and maternal inheritance pattern (see page 56) of mtDNA allows it to be used effectively in forensic identification on poorly preserved postmortem material (e.g., from the bones), where normal nuclear DNA may have long past degraded.

For more thorough coverage of genetic structure concepts, see Trent [5].

Nomenclature, Data Sources, and Online Tools

The Human Genome Project was not the first and is unlikely to be the last large-scale collection of genetic information. Historically, there have been many genetic data collections across a range of species, with a range of nomenclature standards. Indeed, some of the problems still encountered with collating and comparing large historical genetic data sets to gain insight from population level and species comparisons

are due to minor differences in nomenclature and data format.

Given the vast amounts of genetic data already generated and growing, consistency is essential. Fortunately, the collaborative nature of the Human Genome Project drove significant centralization of publicly funded databases and collaboration, with free access to a range of database sets and tools, leading to improved consistency in formats and nomenclature.

The three main publicly available DNA sequence databases, GenBank (NCBI [6], USA), EMBL [7] (Europe), and DDBJ [8] (Japan) formed the International Nucleotide Sequence Database Collaboration [9]. They collaborate and exchange data on a daily basis to ensure public access to up-to-date genetic information, plus useful online tools for interrogating the data. Links to their Websites plus two other useful genome browser online resources, hosted by the UCSC [10] (USA) and the Sanger Institute [11] (UK), are listed in the references.

It should be noted, however, that a minefield of proprietary and privately curated resources also exists. These resources are likely to continue to create compatibility and even ethical challenges from the tsunami of genetic data on its way from new technologies. Auditing of public genetic information resources is also an ongoing process, with large amounts of past genetic data containing ongoing artifacts from historical generation methods, not consistent with current best knowledge. Again, consistency across multiple sources is usually key to the reliability of genetic data.

It is very easy to get lost in the sea of nomenclature conventions, so only some general principles and a few examples will be given here, with links to the main nomenclature bodies for detailed descriptions. A reference summary of genetic nomenclature and database sources, illustrated using mutations from two well-characterized genes and diseases, is provided in Table 3.2. The Atlas of Genetics and Cytogenetics in Oncology and Haematology [12] has a useful, short summary of nomenclature conventions for describing genetic variation.

The HGNC [1] is responsible for overseeing gene nomenclature. Overarching principles they use for gene nomenclature are:

- Try to maintain consistency of names across species.
- Full gene names should be brief and specific and convey character or function (not italicized), e.g., spinal motor neuron protein 1.
- Gene name abbreviations should be italicized, a combination of uppercase letters and numerals, e.g., *SMN1*.
- Protein names should be the same as the gene name but not italicized, e.g., SMN1.

The difference between gene (italics only) and protein name is a subtle but important one that, if adhered to, helps reduce confusion.

The location of a gene can be described by its cytogenetic coordinates, e.g., 7q31-32 (see "Cytogenetics" on page 59), or more accurately by its numerical genomic coordinates, according to a consensus reference human genome, e.g., chromosome (chr)7:117,479,963–117,668,664 for *CFTR*. It is important to cite the reference genome being used as numerical coordinates of genes vary between versions; e.g., *CFTR* starts at chr7:117,120,016 in GRCh37/hg19 and chr7:117,479,963 in GRCh38/hg38.[1] This is because auditing and further annotation continue to refine the inclusion and deletion of data produced from the many genome sequencing projects to date. Given the vast amounts of data involved and the rigorous protocols for reaching consensus, it is important to note that new consensus reference human genomes are only released every few years.

To abbreviate, units of quantity for nucleotide base pair (bp) follow standard SI units, as in computing and other science fields (kilo, mega, giga, tera). It is convenient to drop the "p" from "bp" when given a quantity prefix (i.e., Mbp becomes just Mb). For example:

- 32.1 kb = 32,100 bp
- 3.54 Mb = 3.54 million (3,540,000) bp
- 3.12 Gb = 3.12 billion (3,120,000,000) bp
- chr7:117.48–117.69 Mb

As reference genome coordinate numbers tend to be large, unwieldy, and prone to manual typographic input error, it is acceptable to describe variation in a gene by its HGNC approved name and then the numerical coordinates in the context of the start of that gene's sequence. DNA sequences of interest that need description, in the absence of an HGNC assigned name, may be denoted by their RefSeq label.[2] For example, NC_000007.14 (117470772..117668665) is the RefSeq label for the *CFTR* gene in humans and would be used if it did not already have the name *CFTR*.

Accepted convention for describing variants/mutations is to use the following prefixes, referencing nucleic acid type or protein, to ensure coordinates are consistent:

- c. coding DNA (cDNA)
- g. genomic DNA (gDNA)

[1] GRCh37/hg19 = Genome Reference Consortium human genome build 37 with UCSC genomic annotations version 19 (released in February 2009). As of GRCh38/hg38 (released in December 2013), genome reference build and annotation version number were updated to match each other; however, in some software and reports, you may still see this build referred to as GRCh38/hg20 based on historical numbering conventions.

[2] Hosted by NCBI, the Reference Sequence database is an online collection of curated, non-redundant nucleic acid and protein sequences across species www.ncbi.nlm.nih.gov/refseq; the prefix denotes what type of molecule a RefSeq entry is derived from (e.g. NC = chromosomes, NM = mRNA, NP = protein, NG = genomic).

Table 3.2 Genetic variation/mutation nomenclature and database references

Gene name	Spinal motor neuron protein 1	Cystic fibrosis transmembrane conductance regulator
HGNC gene abbreviation (*italics*)	*SMN1* (HGNC ID:11117)	*CFTR* (HGNC ID:1884)
Protein abbreviation (no italics)	SMN1	CFTR
Gene size (bp)	28,913	188,702
Variant description (cDNA level)	c.836G>T	c.1521_1523delCTT
Variant description (protein level)	p.G279V or p.Gly279Val	p.F508del or p.Phe508del
Cytogenetic gene coordinate	5q13.2	7q31.2
Gene OMIM entry	600354	602421
Disease OMIM entry	253300	219700
Gene genomic coordinates (GRCh38/hg38) UCSC Genome Browser	chr5:70,925,030–70,953,942	chr7:117,479,963–117,668,664
Variant coordinate/s (GRCh38/hg38) ENSEMBL	chr5:70,951,942	chr7:117,559,092–117,559,094
Variant OMIM entry	600354.0005	602421.0001
Variant RefSeq (rs)	rs76163360	rs113993960
Variant dbSNP	76163360	113993960
Clinical significance (ClinVar)	NM_000344.3(SMN1):c.836G>T	NM_000492.3(CFTR):c.1521_1523delCTT
Genetics Home Reference	***SMN1***	***CFTR***

Two examples have been chosen to illustrate nomenclature for genetic mutation. Variants in the *SMN1* and *CFTR* genes (autosomal recessively inherited) have a (autosomal recessively inherited) have a relatively high carrier frequency in many human populations. If both alleles of these genes are adversely mutated, they can result in the conditions spinal muscular atrophy (SMA) and cystic fibrosis (CF), respectively. Entries linked below have specific online database information for that gene or mutation listed as hyperlinks. Tools and information resources associated with genetic analysis can be explored by following the hyperlinks for these genes

URLs:

SMN1—www.genenames.org/cgi-bin/gene_symbol_report?match=SMN1

5q13.2—http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hgFind=omimGeneAcc&position=600354

600354—www.omim.org/entry/600354

253300—www.omim.org/entry/253300

chr5:70,925,030-70,953,942—http://vega.sanger.ac.uk/Homo_sapiens/Gene/Summary?db=core;g=OTTHUMG00000099361;r=5:70925030-70953942

chr5:70,951,942—www.ensembl.org/Homo_sapiens/Location/View?r=5:70951942-70951942

600354.0005—www.omim.org/entry/600354#0005

rs76163360—www.ensembl.org/Homo_sapiens/Variation/Summary?v=rs76163360;toggle_HGVS_names=open

76163360—www.ncbi.nlm.nih.gov/snp/76163360

NM_000344.3(SMN1):c.836G>T—www.ncbi.nlm.nih.gov/clinvar/RCV000009738/

SMN1—http://ghr.nlm.nih.gov/gene/SMN1

CFTR—www.genenames.org/cgi-bin/gene_symbol_report?match=CFTR

7q31.2—http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hgFind=omimGeneAcc&position=602421

602421—www.omim.org/entry/602421?search=CFTR&highlight=cftr

219700—www.omim.org/entry/219700

chr7:117,479,963–117,668,664—http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr7%3A117479963-117668664&hgsid=389964619_QjqACetwCd6XdgESezNnq7o3bzEJ

chr7:117,559,092-117,560,094—www.ensembl.org/Homo_sapiens/Location/View?r=7:117559092-117559094

602421.0001—www.omim.org/entry/602421#0001

rs113993960—www.ensembl.org/Homo_sapiens/Variation/Summary?v=rs113993960;toggle_HGVS_names=open

113993960—www.ncbi.nlm.nih.gov/snp/113993960

NM_000492.3(CFTR):c.1521_1523delCTT—www.ncbi.nlm.nih.gov/clinvar/?term=RCV000058929%20OR%20RCV000007523%20OR%20RCV000007524%20OR%20RCV000119038

CFTR—http://ghr.nlm.nih.gov/gene/CFTR

- m. mitochondrial DNA (mtDNA)
- r. RNA
- p. protein (using amino acid single- or three-letter abbreviation; e.g., G or Gly for glycine)

It is acceptable to describe a variant by referencing its DNA coordinates only, but any protein-level description of a variant must also be accompanied by its equivalent DNA coordinates. As there have been many historical differences in nomenclature and numerical coordinate conventions for describing identical gene changes, conforming to the current HGNC gene name and numerical nucleic acid coordinates for a gene (with or without protein coordinates) minimizes the chance of confusion or error.

Common types of gene variants and some common accepted nomenclature formats are listed as follows (see also mutation descriptions in Table 3.2):

- Single-nucleotide substitution: **c.456G>T** (resulting in non-synonymous amino acid change **p.G152C** or **p.Gly152Cys**)
- Deletion (−3 bp): **c.1521_1523delCTT** (resulting in single amino acid deletion **p.F508del**)
- Insertion (+6 bp between 343 and 344): **c.343_344insCAGTGG** (resulting in two amino acids inserted between arginine at 113 and the amino acid at 114; **p.R113_114insQW** or **p.Arg113_114insGluTrp**)
- Inversion: **c.342_1856inv** (of 1,514 bp fragment)
- Frameshift (downstream stop codon): **p.L125QfsX20** or **p.Lys125Glufsstop20** (lysine at amino acid position 125 is changed to glutamine, with the frameshift extending for 20 amino acids, until a last stop codon)
- Frameshift—from combined deletion (−2 bp) and insertion (+1 bp): **c.2051_2052delAAinsG**

There are many more variations that can be described, but for a fuller explanation of HGNC nomenclature standards, it is best to consult their online resources [1, 13]. Very useful quick reference resources and recommendations for gene nomenclature are also available from the Human Genome Variation Society [14, 15].

It should be noted that shorthand to denote the presence (+) or absence (−) of certain alleles is sometimes used, i.e., homozygous (+/+ or −/−) and heterozygous (+/−), but only when the allele is obvious by context.

Determining if an identified genetic variant is relevant or not is vastly aided by databases that compile genetic diversity, such as dbSNP [16].

Started in 1966 as a collection of known "Mendelian Mutations in Man," a much expanded online version (OMIM) is available today [17]. It provides very useful synopses of recent knowledge and evidence about function, variations, and evidence for pathogenicity in a large number of specific genes (follow links in footnote for Table 3.2 for examples of listings for *CFTR* and *SMN1* genes).

The database of Genotypes and Phenotypes [18] is a useful resource for trying to correlate clinical manifestation with genotypic change. The Human Phenotype Ontology [19], initially formed by mining information from OMIM, attempts to standardize vocabulary for phenotypic classification. Its rigid hierarchical nature, necessary for its aims, is a little user-unfriendly. PhenomicDB [20] attempts to compare phenotypes across many species with much genotypic information logged. Regions of genes highly conserved across species tend to indicate an evolutionary tendency to conservation for survival, therefore a higher likelihood of pathogenicity in variants.

Genetic problems often express themselves at enzyme or protein level with the cited online resources useful [21–23].

The new field of epigenetics (discussed on page 57) has spawned its own nomenclature challenges [24].

Although somatic mutations are only very briefly covered in this chapter (see later text), COSMIC [25] is a useful tool for this area.

In cytogenetics, the accepted nomenclature bible *An International System for Human Cytogenetic Nomenclature* [26] is unfortunately only available as hard copy and not available online. As karyotyping is not the main impetus of this section, an explanation of the basics of cytogenetic nomenclature, with a few simple examples, is given in the section on cytogenetics (page 59).

A range of publicly funded resources aimed at demystifying genetic information for lay audiences exist. They can be a good starting point, even for medical professionals (e.g., Genetics Home Reference [27]). Devoted education sections of professional journals or research bodies are also a good resource, summarizing complex scientific concepts into a more digestible form (e.g., Nature's Scitable [28], NIH National Human Genome Research Institute fact sheets [29]).

Inheritance

Knowledge of modes of inheritance is essential to understanding genetic disease processes. Most of our attributes, good and bad, are directly linked to inheritance from our ancestors. Austrian monk, Gregor Mendel, is credited as the first to describe the process of genetic inheritance, from experiments conducted in the mid-1800s, many years prior to the elucidation of DNA as the carrier of the genetic blueprint. Although not coining the term himself, he was the first to outline the concept of an "**allele**" to describe alternative forms of the same gene or genetic element (**genotype**). As humans normally have two copies of the same gene (one inherited from each parent), it is the expression and interplay of these two alleles that determine expression of **traits**, i.e., characteristics. **Phenotype** refers to the trait/s actually expressed physiologically and may diverge from that expected for a certain genotype.

Mendel's experiments with breeding garden peas and assessing mainly binary traits (e.g., color) led to three laws:

- Law of segregation: when gametes form, they only retain one copy of a gene for a given location (one **allele**).
- Law of independent assortment: genes can segregate independently when gametes are formed (**recombination**).
- Law of dominance: some alleles are **dominant** (express even if another allele is present) and some are **recessive** (only express if both alleles are recessive). The law of dominance underpins what is referred to today as "**Mendelian inheritance**" or a "**Mendelian trait**"; i.e., inheritance follows an autosomal **dominant** or autosomal **recessive** pattern in a single gene.

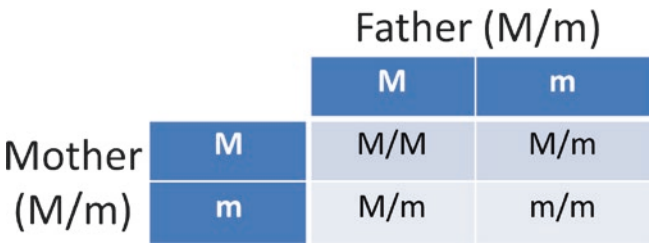


Fig. 3.6 Inheritance pattern from heterozygous parents. "Punnett square" indicating inheritance of autosomal recessive (m) or dominant (M) allele from heterozygous carrier parents

A recently compiled summary of listings on OMIM indicates 94 % autosomal, 6 % X-linked, 0.3 % Y-linked, and 0.3 % mtDNA diseases [5].

Conventionally, a dominant Mendelian allele is represented by a capitalized letter (M) and recessive allele by a lowercase letter (m). There are then three possibilities of segregation depending on what alleles the parents have: *M*/*m* (**heterozygous**) and *M*/*M* or *m*/*m* (**homozygous**) (see "Punnett square" box in Fig. 3.6). A dominant allele will express if present, whether a recessive allele is present or not (M/M or M/m). A recessive allele will only be expressed in the phenotype if both alleles are recessive (m/m).

Autosomal recessive traits are inherited in a horizontal manner (see Fig. 3.7a). In the offspring of heterozygous (carrier) parents, there is a 25 % chance of autosomal recessive allele being expressed and 50 % chance of being a carrier of the recessive allele (not expressed).

Cystic fibrosis (CF) is an example of autosomal recessively inherited disease (*CFTR* gene), most frequently homozygous for the most common mutation (F508del/F508del; c.1521_1523delCTT/c.1521_1523delCTT). However, CF also demonstrates the concept of a **compound heterozygote**, when two different disease-associated recessive alleles in the same gene are expressed (e.g., F508del/G542X; c.1521_1523delCTT/c.1624G>T), resulting in a disease phenotype.

Closely related individuals have a higher chance of carrying similar DNA, as they have closer common ancestors. Therefore, **consanguinity** increases the chance of autosomal recessive traits being expressed; i.e., the chance of alleles from parents being the same is increased the more closely they are related genetically. The **coefficient of inbreeding** (**f**) measures the theoretical level of homozygosity based on **pedigree**, with first cousins expected to share one eighth of their DNA, therefore having approximately 12.5 % homozygosity ($f=0.125$). The prevalence of certain alleles also differs between ethnic groups, again due to effects of closer common ancestors.

Autosomal dominant traits are inherited in a vertical manner, with a 50 % chance of being passed onto offspring (Fig. 3.7b). There may, however, be a range (from minor to severe) of disease traits expressed in different individuals with the same dominant allele (**variable expressivity**). Some alleles may be present, but not express themselves in all individuals; **penetrance** refers to the percentage of individuals expressing the phenotype associated with a specific allele by a certain age (e.g., evidence of autosomal dominant hypertrophic cardiomyopathy is dependent on age and differs even within families). For a specific allele, penetrance refers to the chance of a phenotype being present (or not). In contrast, expressivity refers to the severity of traits expressed, implying that there is a level of phenotypic expression present, however minor it may be.

Pleiotropy (literally "affecting many") describes where a single allele manifests phenotypically in multiple, apparently unrelated traits. Modulation of these traits may be impacted by environmental and other factors. Monozygotic twins demonstrate this concept well. Despite identical genotypes (i.e., an identical complement of alleles), monozygotic twins can express traits differently—i.e., have **discordant phenotypes**. This indicates that there are factors other than genotype that can affect phenotype (see epigenetics on page 57).

Haplotype refers to a subset of the genotype, usually of alleles that tend to be inherited together and frequently from one parent. The concept of haplotype is important in historical methods used to isolate candidate disease genes through **linkage analysis** of affected individuals and families (e.g., *CFTR* gene in cystic fibrosis). This method relies on non-disease marker genes in close proximity to a disease gene frequently being inherited together, acting like a flag to the disease gene. Sometimes genes in close proximity (**contiguous**) may all be affected together by relatively large DNA changes, leading to complex phenotypes that are a combination of the multiple allele changes (e.g., 11p14 deletion causing aniridia and increased risk of Wilms tumor). **Hemizygous** refers to the loss of one of a pair of chromosomes, either whole or in part. **Haploinsufficiency** (reduction of relative gene expression from loss of one allele) can result from a reduction in gene dosage in hemizygosity and lead to disease (e.g., 7q11.23 deletion of 26 genes in Williams syndrome).

Sex-linked inheritance follows an oblique inheritance pattern associated with segregation of the X and, very rarely, the Y chromosome (Fig. 3.7c). Males are referred to as **hemizygous** for the entire X chromosome, as although diploid for the autosomes they have only one copy of the X chromosome. Fabry disease and hemophilia (A and B) are **X-linked disorders**, expressed in males in a hemizygous manner. Fabry disease may also present in the phenotype of heterozygous females to varying degrees, through the process of **X-inactivation** (**lyonization**). This is the process whereby one X chromosome in each cell is randomly made transcriptionally inactive through chromatin structure changes at the time of embryo development (see epigenetics

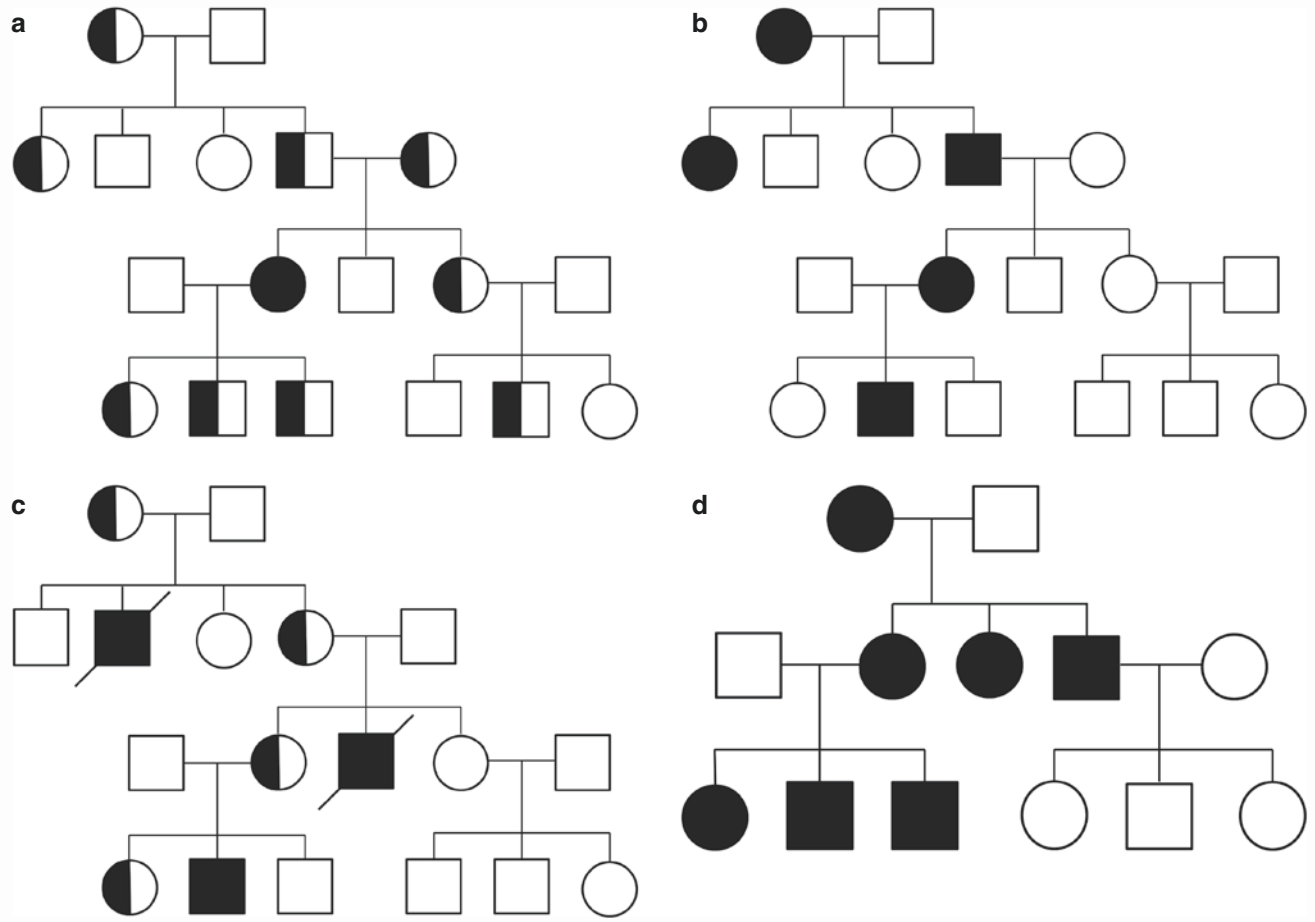


Fig. 3.7 Inheritance pattern genograms (pedigree). (**a**) Autosomal recessive (AR): if both parents are carriers of an AR mutation, there is a 25 % chance of their child being homozygous for the mutation and 50 % chance of them being a carrier. All children from one homozygous and one non-carrier parent will be carriers of an AR mutation (*bottom left*). (**b**) Autosomal dominant (*AD*): if either parent is affected, there is a 50 % chance that their child will be affected. Age of onset and severity of disease will be dependent on penetrance and expressivity, respectively. (**c**) X-linked (*XL*): a mutation is passed on through an X chromosome. As females have two X chromosomes, a healthy allele on one X chromosome most often compensates for a mutation on the other X chromosome. X-linked conditions most often affect males, as they only have one copy of the X chromosome, with no other allele to compensate, leading to disease if their only X chromosome contains a pathogenic mutation. Sons of carrier mothers or affected fathers have a 50 % chance of being affected. Daughters of carrier mothers have a 50 % chance of being a carrier of an X-linked condition. (**d**) Mitochondrial (*mtDNA*): all children of an affected mother will carry a mitochondrial DNA mutation, as this sub-organelle DNA is only inherited from the mother. Affected males do not pass on mitochondrial DNA mutations to their children, as mtDNA is normally only inherited from the mothers (female gamete). **Key** *Square*: male. *Circle*: female. *Full-shading*: affected. *Half-shading*: carrier. *No shading*: unaffected. *Diagonal line*: deceased

on page 57). **Sex-determining region Y** (**SRY**) **protein** on the Y chromosome is responsible for the initiation of male sex determination, and faults in its expression can be responsible for aberrations between sex phenotype and genotype.

Most traits are thought to be under more complex control than Mendelian inheritance, via **incomplete dominance** (both alleles expressed to some degree, with the phenotype a combination of their expression, e.g., sickle cell trait that is milder than the homozygous [*HbS*/*HbS*] sickle cell anemia), **codominance** (both alleles expressed in the phenotype, e.g., ABO blood grouping), or **digenic/polygenic** (influenced by two or more genes, e.g., autosomal recessive retinitis pigmentosa, autosomal recessive hearing loss). Mitochondrial disease follows a pattern of **maternal inheritance** only (Fig. 3.7d), as mitochondrial DNA (mtDNA) in a zygote is derived exclusively from the maternal oocyte. Therefore, all children from the same mother can have the same mitochondrial-derived trait; however, only daughters can pass it on to their offspring.

A **genogram** (family tree or **pedigree**) is a useful method for visualizing inheritance and is often used to elicit the likely segregation pattern (Fig. 3.7). This can be a useful aid in refining differential diagnoses and genetic tests to be performed.

Epigenetics

Epigenetics is a relatively new field that has generated a wealth of interest, especially in its implications for genetic disease and testing. The prefix *epi* (Greek for "over" or "above") infers a meaning of genotypic effect over and above that performed by the genome; however, its definition continues to be debated, particularly with regard to mechanisms that are not heritable. It is generally agreed that **epigenetics** refers to modulation of gene activity or expression without modification to gene sequence. The term **epigenome** is used to describe the complement of all epigenetic effects. The NIH Roadmap Epigenomics Project Consortium Project [22] includes both heritable and non-heritable mechanisms in its definition, agreed to here for the purposes of discussion.

The starkest demonstration of epigenetic mechanisms is when monozygotic twins with identical genotypes express differences in phenotype, by the presence or absence of disease [30]. The depth of knowledge of this mechanism of genetic modulation and its impact on all manner of disease is still relatively new but is increasingly finding its way into genetic diagnostics. Like dark matter in physics, epigenetics may well turn out to be the previously hidden mechanism behind a range of phenotypes not explained using classical genetic models. The hope is that it will become an important aid in determining why one person gets a disease and another of similar genotype remains unscathed.

Genomic imprinting, where an allele is completely silenced based on its parental origin, is an epigenetic phenomenon responsible for diseases such as Beckwith-Wiedemann syndrome, Prader-Willi syndrome (paternal inheritance), and Angelman syndrome (maternal inheritance). Epigenetic phenomena also underlie the process of X-inactivation (for review, see [31, 32]).

While further types of epigenetic regulation are likely to be discovered, the following mechanisms (all post-translational) are already known to be the basis of several epigenetic phenomena, with relevance in disease. This whole field is currently one of the most active areas in biomedical research.

Nucleosome Position

DNA is packaged into the nucleus wrapped around **histone** proteins to form **nucleosomes**, making up the majority of the **chromatin** complex. Changes in the position of nucleosomes in the chromatin structure can affect gene transcription mechanisms by altering proximity and/or access to transcription start sites.

Histone Modification

Modification of histone N-terminal tails by methylation, phosphorylation, acetylation, ubiquitination, sumoylation, ribosylation, or citrullination can alter the initiation of transcription of a gene. Like nucleosome positioning, it can act by altering the chromatin structure, modifying either positively or negatively the ability for transcription to initiate at specific sites. Histone modification has also demonstrated wider reach, able to affect DNA repair and replication, plus alternative splicing mechanisms.

CpG Methylation

Probably the most widely known and tested form of epigenetic modification, **methylation** of specific cytosine nucleotides can repress gene expression by inhibiting transcription factor binding and enhancing recruitment of chromatin co-repressors. Cytosine nucleotides adjacent to a guanine (commonly referred to as **CpG** for **c**ytosine joined by a **p**hosphodiester bond to adjacent **g**uanine) are the targets for this methylation via DNA methyltransferase (DNMT) enzymes. This tends to happen in CpG-rich regions (called **CpG islands**), which frequently occur near to 5′ gene promoter regions. Their effect is to repress transcription, effectively silencing a gene. The equivalent of single-nucleotide polymorphisms (SNPs) for the genome, **methylation variable positions** (**MVPs**) are sites that show common variability in their effect on epigenetic regulation. Epigenomic maps of such information are continuing to evolve, and the term **methylome** is now used to describe the entire complement of methylated CpG sequences.

Non-coding RNA

Surprisingly, only 20 % of RNA (mRNA) is translated into protein. The question remains then: What might be the function of the remaining 80 % of RNA transcripts (termed **non-coding RNA**; **ncRNA**)? At least some ncRNAs are involved in epigenetic forms of regulation, through what is termed **RNA interference** (**RNAi**). The short (20–25 bp) double-stranded molecules of **microRNA** (**miRNA** not to be confused with messenger RNA [mRNA]) and **silencing RNA** (**siRNA**) have different but overlapping roles. Both act by directly binding to mRNA molecules, miRNA less specifically than siRNA. siRNA actively degrades already transcribed mRNA through the actions of the enzyme Dicer and protein complex RISC (see [33] for excellent animation of the process). miRNA acts to indirectly prevent translation to protein just by virtue of it binding to the 3′ untranslated

region of an mRNA molecule, but it can also utilize the same degradation pathway of Dicer and RISC as siRNA.

Although an arbitrary value to distinguish them from the shorter ncRNAs, **long non-coding RNAs** (**lncRNAs**) are at least 200 bp but frequently much larger [34]. They work in a variety of ways, but an example is the very well-characterized X-active specific transcript (XIST). XIST is a 17 kb lncRNA responsible for mediating **X-inactivation** by effectively coating the X chromosome it is transcribed from, rendering it inactive. lncRNAdb is a database focusing on lncRNAs with experimentally characterized function [35].

Other ncRNAs involved in epigenetic processes but beyond the scope of this chapter include ribozymes ("gene shears"), Piwi-interacting (piRNA), small nuclear (snRNA), small nucleolar (snoRNA), and transcription initiation (tiRNA) RNA.

Exogenous manipulation and monitoring of ncRNAs, especially miRNA and siRNA, have spawned a whole new range of potential diagnostic and therapeutic possibilities—although still predominantly in the research phase.

It should be noted that the aforementioned epigenetic mechanisms are often interactive, not necessarily acting in isolation, each able to up- and downregulate the likelihood of one of the others coming into play and acting in concert to modify chromatin structure and/or gene expression. X-inactivation is an example of several of these mechanisms working in tandem for epigenetic regulation.

The International Human Epigenome Consortium (IHEC) launched the Human Epigenome Project [36, 37] in 2010, aiming to "decipher at least 1,000 epigenomes within the next 7–10 years," to determine epigenomic impact on "… key cellular status relevant to health and disease" [38]. GenomeRNAi is a database compiling phenotypes resulting from RNA interference [39].

It was previously thought that mitochondrial DNA (**mtDNA**) was only capable of modulating important phenotypic effects by acting on the nuclear DNA epigenome. However, recent emerging evidence suggests that mtDNA itself may be able to be directly epigenetically regulated, with many fascinating implications.

Somatic Mutations

The genotype of subsets of cells and tissues may change throughout life from normal wear and tear, accumulation of errors through normal regulation and repair, or exogenous factors, such as adverse environmental exposures (e.g., radiation, toxins). Cancers, on the whole, develop in this manner, first localizing abnormalities to cell subtypes, tissues, and regions and then spreading through metastasis. Genetics in this area would require another whole chapter to discuss, but it is just highlighted here in order to flag the rare occasions where tumors can develop *in utero* and be the obvious cause of pathology. The genetic tests for somatic cancer are obviously indicated at these times.

Genetic Testing

In an attempt to simplify the concepts of many of the tests presented below, a book analogy is used where possible.

> If one book is a gene, then the genome is a whole library, shelves of books are equivalent to the chromosomes, and individual letters on the page are the single nucleotides of DNA.
>
> Expanding the analogy further, collaborations such as the 1000 Genomes Project are like an international congress of libraries, pooling all of their available books, information, and resources together.

Sampling

Genetic testing requires isolation of nucleic acid (DNA or RNA) (for a quick reference summary, see Table 3.3). RNA degrades much more rapidly than DNA and therefore requires more careful handling and extraction. In general, the most reliable and most frequently used sample type for genetic testing is blood transported at room temperature in an EDTA tube. Cord blood can be a useful source for testing in the early neonatal period. If blood is not available (e.g., postmortem cases), then heart, lung, and other tissues may be used directly to isolate DNA (preferably not the liver as its protein- and enzyme-rich composition tends to hamper good nucleic acid isolation). The skin is very robust, but lung and other tissue may also be used to culture cells from which DNA can be isolated. This tissue is best provided fresh on its own in a sterile sample container or in culture media (e.g., RPMI) or normal saline, stored at room temperature for short periods or 4–8 °C (not frozen) for up to a few days.

Amniocentesis (amniocytes) and chorionic villous (placenta) sampling also rely on cell culture to derive cells for DNA isolation or karyotyping. Given the small amount of material and their relative scarcity, processing is best performed immediately; therefore, forewarning the laboratory about these procedures is essential. Cytogenetics uses blood in lithium heparin or sodium heparin tubes for isolation of peripheral blood lymphocytes (PBLs) to culture for isolation of chromosomes.

It is possible to isolate DNA from formalin-fixed, paraffin-embedded (FFPE) tissue, but the process of fixation causes significant degradation to nucleic acids. DNA extraction can be attempted on these samples, but quality and quantity isolated are inconsistent, with a high failure rate, making this

Table 3.3 General guidelines for obtaining DNA samples

Test	Tissue	Target	Collection vessel	Transport/storage
Karyotype/FISH	Blood	Culture PBLs	Li Hep/Na Hep	Room temp. <72 h
	CVS	Placenta	Sample jar with sterile culture media	Room temp. if immediate processing, 4 °C if >48 h until processing
	Amniotic fluid	Amniocytes	Plain sterile tubes	Room temp. <48 h
All other nucleic acid-based testing	Blood	PBL DNA	K-EDTA/Na-EDTA	Room temp. <72 h
	Heart, lung, other tissues	Direct DNA isolation	In sterile sample jar	4 °C or −20 °C (do not freeze if also used for culture)
	Skin, lung, heart, liver, other tissues	Cultured cells used for DNA isolation	In sterile sample jar in RPMI, normal saline, or tissue on its own	Room temp. or 4 °C < 48 h
	FFPE tissue (tumor)	DNA isolation	Dewaxed on slide	Room temp. in slide box

PBL peripheral blood lymphocyte, *Li Hep* lithium heparin blood collection tube, *Na Hep* sodium heparin blood collection tube, *Room temp*. room temperature, *CVS* chorionic villus sample, *RPMI* a type of cell culture media, *FFPE* formalin-fixed, paraffin-embedded, *h* hours, *K-EDTA* Potassium EDTA blood collection tube, *Na-EDTA* Sodium EDTA blood collection tube

not a preferred option for germline genetic testing. For somatic genetic testing, where the majority of the FFPE sample is tumor DNA, extraction can be more useful and consistent. If used, FFPE samples should be provided dewaxed on original slides, with tumor-rich regions marked in some way.

Maternal blood in EDTA tubes is used to isolate circulating free DNA from plasma (see NIPT; page 79).

It should be noted that while theoretically all of our cells should have the same genotype, **mosaicism** (genotypes divergent between cells in the same individual) can occur. Any isolated DNA will be representative of the cell or tissue type it is derived from, which may not always be representative of the genotype of all cells in the body (e.g., **placental mosaicism**).

The aforementioned are general guidelines only, and laboratory resources or staff should be consulted to determine what tests are available and the most suitable sampling, storage, and transport methods for your local service.

Complementarity: The Basis of Genetic Testing

The machinery of DNA replication (detailed previously) underpins the mechanism behind almost all genetic testing, other than karyotyping. Binding of a nucleotide to its complementary nucleotide in an antiparallel, mirrorlike fashion gives the structure of DNA many advantages in terms of fidelity for replication and repair. Genetic testing relies on the fact that a nucleotide sequence AGCTGGCT will only bind to its complementary sequence TCGACCGA (UGCTGGCT if RNA) and is the basis of the incredible precision possible with genetic testing. Harnessing the power of enzymes involved in the fundamental processes of DNA replication also allows very small amounts of starting material to be amplified into sufficient quantities for a range of different genetic tests. Cytogenetics is the exception; staining DNA with dye.

Cytogenetics

Cytogenetics is the study of chromosomes, with their number and characteristics assessed to produce a **karyotype** (*karyon* from Greek for nucleus). By visualizing banding patterns on stained chromosomes, DNA can be analyzed at a gross level, with changes detectable in the 5–10 Mb range (~400 band resolution).

> If DNA sequencing is analogous to reading single letters of a book, then karyotyping is like looking at the shelves of a library through a telescope, from a marked distance away.

In cytogenetics, it is important to be aware of two of the phases of mitosis. Approximately 90 % of a cell's life cycle happens in **interphase**, where chromosomes are highly condensed in the nucleus. As most cells are already likely to be in interphase before cell culture begins, the lead time to being able to harvest interphase cells can be as short as 24–72 hours. In **metaphase**, chromosomes align along the equator of the cell guided by microtubules. It is at this time that chromosomes are most easily visualized, which is therefore the preferred state for karyotyping. The disadvantage of examining metaphase cells though is that the process can take considerably longer than preparation of interphase cells (usually 1 week but often longer for slow growing cells or other problems requiring repeat culture).

The general technique for karyotyping comprises the following steps:

- Use a sample of blood to start cell culture of peripheral blood lymphocytes (PBLs).
- Stimulate cultured cells with phytohemagglutinin to force them toward metaphase.

- Add colchicine to arrest cells at metaphase.
- Use a hypotonic solution to swell and burst open cells, to release metaphase chromosomes, and to enhance their spreading, i.e., remaining in single clumps from single cells, but separated sufficiently from each other so the individual chromosomes can be visualized under a microscope.
- Drop spread chromosomes onto slides and then fix and stain to visualize banding patterns. G-banding uses the most common stain Giemsa (methylene blue, eosin, and azure B).
- Microscopic examination.

Each chromosome has a consistent and well-characterized banding pattern, centromere location, and length allowing it to be identified and classified. **Heterochromatin** refers to the dark bands from densely packed DNA. **Euchromatin** is the lighter regions, gene-rich, and more accessible for active transcription. Scoring individual chromosomes from a number of cells on a slide allows the determination of gross changes that may indicate **aneuploidy** (anomalies in the total number or character of chromosomes).

Chromosomes are arranged in pairs of **sister chromatids** connected by a **centromere**. The centromere creates a division into two arms for each chromatid, with the shorter arm labeled **p** (from the French "petit") and the longer arm labeled **q** (as it follows p in the alphabet). Location is classified by sequential numbering starting from the centromere and moving outward (i.e., proximal to distal) on both arms. The first two numbers are region and band, respectively, (e.g., q23 is region 2, band 3). The region and band should always be stated as single numbers (i.e., for the previous example two-three, not 23) unless you want to raise the ire of a cytogeneticist. The centromere is the start of region 1, and subbands follow a decimal point after the region and band number; e.g., 13q23.1 is subband 1 of band 3, region 2 distal from the centromere on the q (long) arm of chromosome 13.

A karyotype is reported by a numerical value of the number of chromosomes (normal in humans is 23 pairs = 46), then sex chromosomes, and then, if present, any aneuploidy. Parentheses identify the type of rearrangement, a semicolon separates alterations in two or more chromosomes, and a tilde (~) is used to show uncertainty in the location. The total number of cells counted is indicated in square parentheses at the end. Strict nomenclature guidelines are provided by the International Standing Committee on Human Cytogenetic Nomenclature [26].

A **normal karyotype** is **46,XX** (female) and **46,XY** (male).

Examples of a female trisomy 13 (47,XX,+13) and a female triploid karyotype (69,XXX) are given in Figs. 3.8 and 3.9, respectively.

The main types of aneuploidy are **duplication**, **deletion**, **translocation**, **inversion**, **isochromosome**, **ring chromosome**, and **uniparental disomy** (**UPD**). **Terminal** and **interstitial** changes (usually deletions or duplications) refer to those near the ends and within the internal part of a chromosome, respectively. Table 3.4 gives examples of these types of aneuploidy with an example karyotype and common disease name.

Mosaicism refers to cases where there are cells with more than one karyotype in the same individual. There are many causes, especially aging, but all mosaic karyotypes are generated from only one zygote (Table 3.4). **Placental mosaicism** can be a cause for apparent trisomy (in cells from the placenta only) that is not present in the fetus.

Although very rare, **chimerism** is where more than one karyotype exists in the same individual, originating from separate individual zygotes (Table 3.4). This occurs after successful bone marrow or other tissue transplants but prenatally is usually the result of early embryonic twin-twin fusions resulting in a dual karyotype singleton.

Non-invasive prenatal testing (NIPT; see page 79) is currently making large inroads into replacing karyotyping for prenatal screening. However, karyotyping remains the gold standard and is still used to confirm positive NIPT results.

- **Traditional karyotyping** is a good test for detecting trisomies (13, 18, 21) and often indicated for multiple or suspicious miscarriages.

Fluorescent *In Situ* Hybridization

Fluorescent *in situ* hybridization (**FISH**) is used in cytogenetics as an alternative, as well as adjunct to karyotyping. As it can be used on interphase cells, it allows for more rapid detection of suspected aneuploidy. It can also be used to confirm or further characterize karyotype results. It relies on fluorescently labeled DNA probes (10–100 kb) that hybridize to complementary regions of DNA on chromosomes. Tens of thousands of commercial and in-house probes exist, many generated from the sequencing techniques employed in early parts of the Human Genome Project. Usually, only a small subset is used for rapid assessment or confirmation according to the suspected aneuploidy. FISH has the advantage of a relatively quick turnaround time (approximately 48–72 hours from sample receipt).

The technique is similar to most nucleic acid hybridization techniques, i.e., heat to denature DNA into single strands, followed by the addition of a labeled single-stranded DNA probe that will bind to its complementary sequence. For FISH, this occurs in the fixed tissue (cells) on a slide, hence the "*in situ*" component of its name.

To use the book analogy, FISH is analogous to locating a fluorescently painted part of a shelf (when using a telescope to look through a library window from a marked distance away).

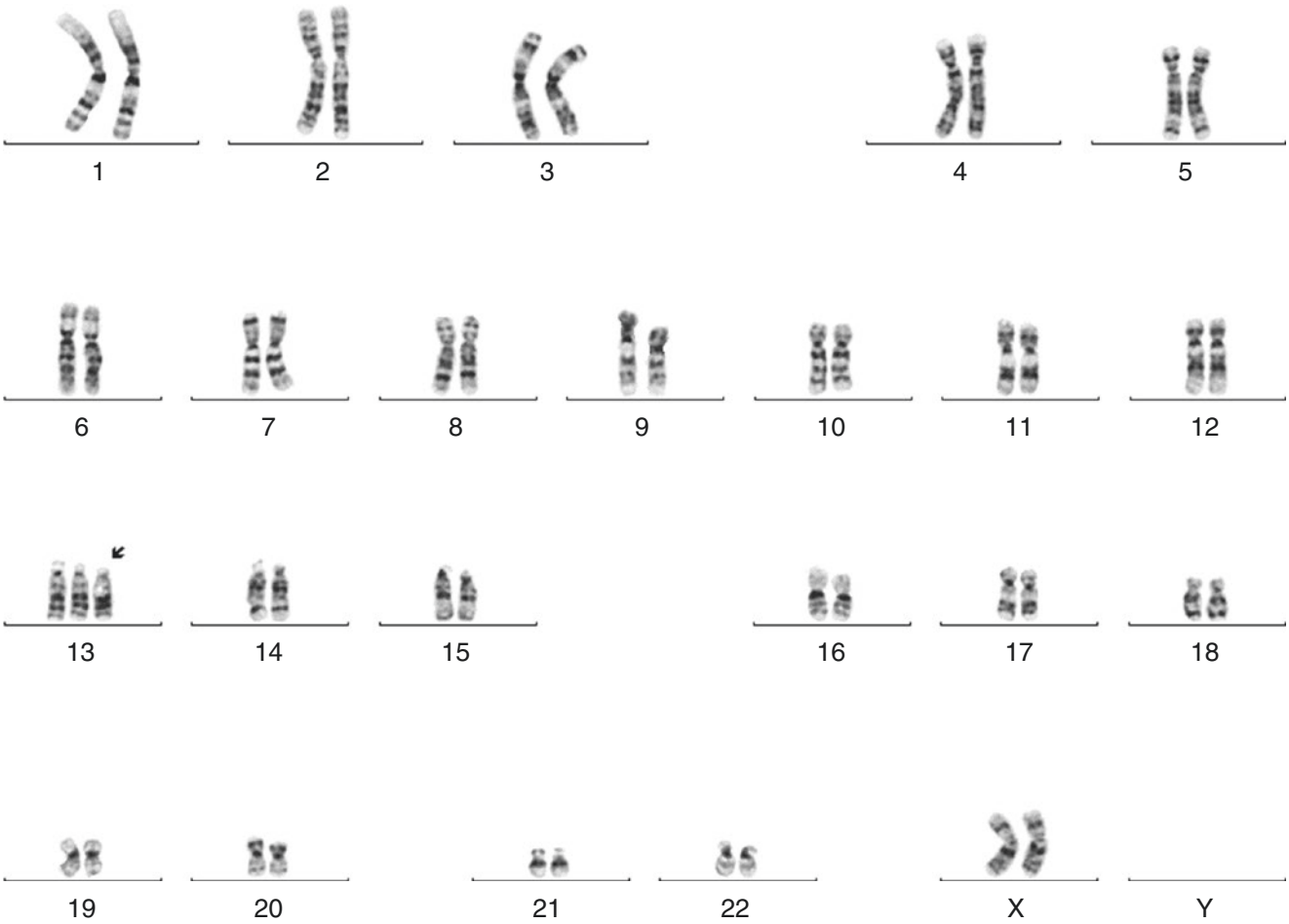


Fig. 3.8 Trisomy 13 (Patau syndrome) karyotype. One extra copy of chromosome 13 (*arrow*) indicating a female with trisomy 13 (karyotype notation = 47,XX,+13) (Figure courtesy of Ms. R. Hutchinson, SA Pathology, Australia)

A range of different FISH probes exists, allowing different lengths, parts, and characteristics of chromosomes to be visualized (e.g., translocation, centromere, subtelomere, fusion, breakpoint, and painting probes). The latter use multiple probes to color-code all chromosome pairs different colors in the one reaction. Simple examples of trisomy 21 and sex determination by FISH are shown in Figs. 3.10 and 3.11, respectively.

A FISH result is denoted by "**nuc ish**" (for nuclear *in situ* hybridization) for the karyotype, with probe name in parentheses and cell number counted in square parentheses following, e.g., nuc ish(D21S259/D21S341/D21S342)x3 [200/200]. Often the FISH result is reported first verbally, but usually karyotyping is also commenced in parallel and reported later with a metaphase FISH result for confirmation. A standard karyotype is listed first, followed by the FISH result (see Table 3.4).

The principles behind FISH also form the basis of microarray hybridization techniques (see page 73).

- **FISH** is a good test for rapid assessment of trisomies, frequently used for fetuses rapidly approaching the cutoff age for termination. FISH is useful when targeting particular areas of the genetic code.

Automated DNA Sequencing

Named after its inventor, dual Nobel laureate in chemistry, Frederick Sanger, dideoxynucleotide (**Sanger**) **sequencing** was one of many systems he trialed, outlasting them and other competitors. Until very recently, it has been the mainstay of DNA sequencing.

To use the book analogy, DNA sequencing is equivalent to reading a book from the first page to the last page and then reading it backward from the last page to the first page, looking for spelling errors in individual words.

Sanger sequencing uses the following steps to replicate the targeted region into many individual fragment chains differing by single nucleotides in their size. Separating them by size gives a ladder- or barcode-like pattern indicating the DNA sequence (Fig. 3.12a):

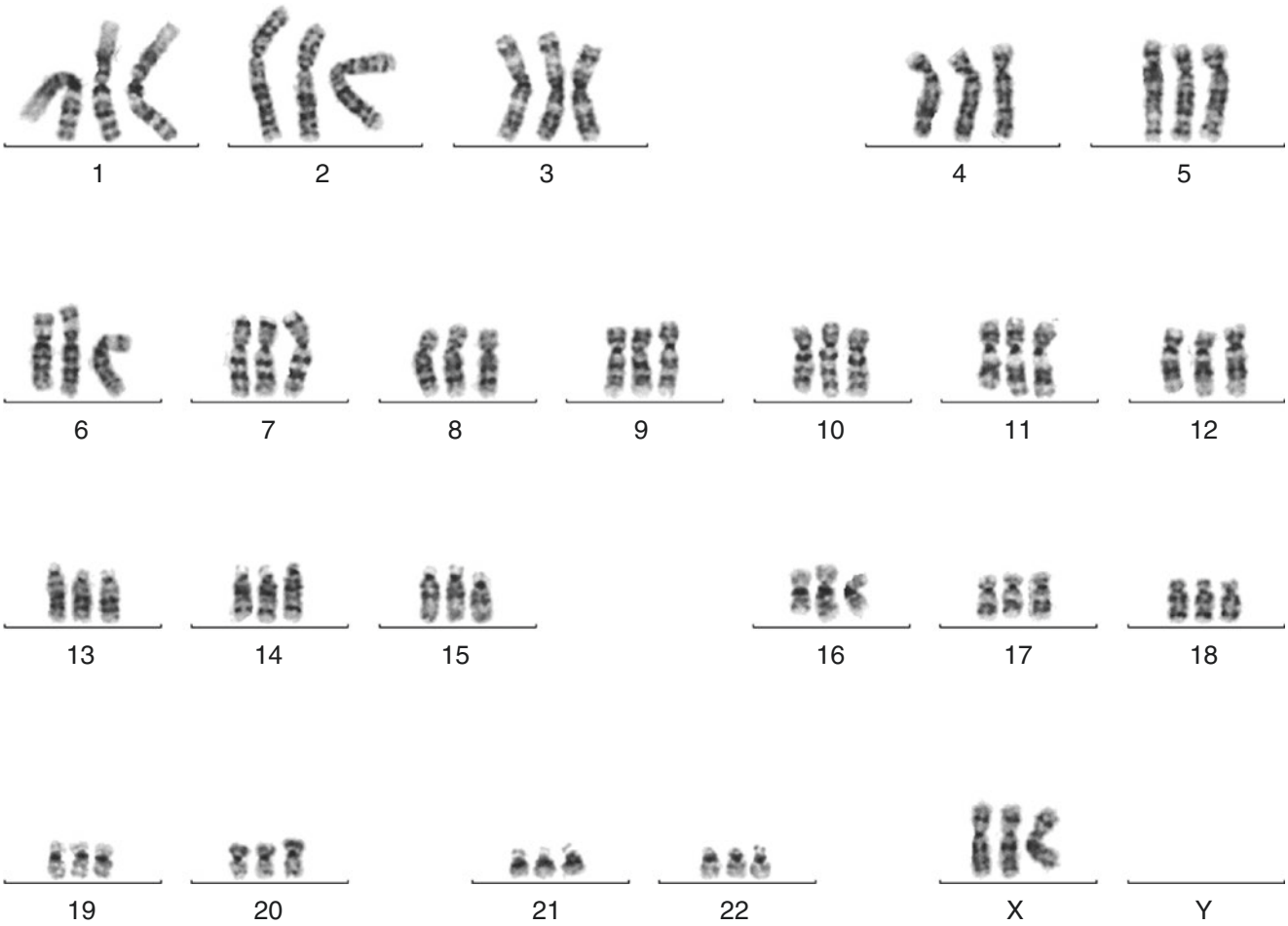


Fig. 3.9 Triploid karyotype. Three copies (3n) of each chromosome in a female (karyotype notation = 69,XXX) (Figure courtesy of Ms. R. Hutchinson, SA Pathology, Australia)

Table 3.4 Karyotype—examples of cytogenetic abnormalities and nomenclature

Duplication: replication of all or part of a chromosome		
47,XY,+21	Trisomy 21 (Down syndrome)	One extra chr21 ➪ total 47
47,XX,+18	Trisomy 18 (Edwards syndrome)	One extra chr18 ➪ total 47
47,XXY	Klinefelter syndrome	One extra sex chromosome ➪ XXY (total 47)
46,XX,dup(8)(p22p21.1)	Partial trisomy 8	Duplication and inversion of part of chr8 between region 2, band 1, subband 1, and band 2 (NB: inversion in duplications is indicated by the reversal of band number order, i.e., 22 before 21)
Deletion: loss of all or part of a chromosome		
45,X	Turner syndrome	One missing X chromosome
46,XX,del(5)(p13)	Cri du chat syndrome	Deletion of short arm (p) of one chr5 from region 1, band 3, to the subtelomere of the short arm
46,XX,del(1)(p36.3)	1p36 deletion syndrome	Deletion of short arm (p) of one chr1 from region 3, band 6, subband 3, to the subtelomere of the short arm
Translocation: relocation of all or part of a chromosome so that it is incorporated into another chromosome		
46,XY,t(9;22)(q34;q11.2)	Philadelphia chromosome	Translocation between long arms (q) of chr9 and 22
47,XX,+der(22),t(11;22)(q23;q11)	Miscarriage	Translocation between long arms (q) of chr11 and 22, with derivative of chr22 producing trisomy

Table 3.4 (continued)

Translocation: relocation of all or part of a chromosome so that it is incorporated into another chromosome		
46,X,der(X),t(X;Y)(q28;p11.31)	Ambiguous genitalia	Translocation (Robertsonian, i.e., reciprocal) between the long (q) and short (p) arms of chrX and chrY, respectively
Inversion: part of a chromosome that has reversed its direction 180°, so that it is oriented in the opposite direction on the chromosome to normal		
46,XY,inv(7)(p22;q22)	Some cases of fetal demise	Inversion between region 2 band 2 on short arm (p) and region 2 band 2 of long arm (q) of chr7
Isochromosome: a chromosome that has lost one of its arms and replaced it with a copy of the same arm (i.e., p-p or q-q)		
46,XX,i(18)(q10)	Isochromosome 18q syndrome	One chr18 has two long arms (q); the breakpoint is assigned the centromere location q10 (region 1 band 0)
Ring chromosome: arms of a chromosome have fused together in a ring shape		
46,XX, r(15)	Ring chromosome 15 syndrome	One chr15 has fused short and long arms into a ring
Uniparental disomy: both chromosomes of a pair, or parts of them, are derived from the same parent		
46,XY,upd(16)mat	Associated with some cases of IUGR	Both of the chr16 pair are derived from the mother
Mosaic: more than one karyotype in the same individual (derived from one zygote)		
mos 45,X/46,XX	Turner syndrome mosaicism	Two subsets of cells from the same zygote with different karyotypes
mos 46,XX^{SRY+}/45,X^{SRY+}	Ovotesticular disorder of sexual development (OT-DSD)	Two subsets of cells, both containing male sex determining region (SRY), Turner syndrome on the X chromosome mosaic
Chimera: more than one karyotype in the same individual (derived from more than one zygote)		
chi 46,XX/46,XY	Female/male chimera	Two subsets of cells from two zygotes with different karyotypes
FISH: for rapid analysis (interphase) or confirmation (metaphase) of aneuploidy		
Interphase (not usually reported, other than verbally)		
nuc ish(D21S259/D21S341/D21S342)x3	Trisomy 21	Three fluorescent signals are detected for chr21 in the same cell
Metaphase (clinical FISH, reported with karyotype)		
46,XX.ish del(22)(q11.2q11.2)(TUPLE1-)	DiGeorge (22q11.2 deletion) syndrome	Absence of signal on the long arm (q) for region 1, band 1, and subband 1 of chr22, confirmed by FISH

NB: Cytogeneticists will not normally refer to a location according to region, band, or subband; it is used here only to illustrate the systematic approach of defining chromosomal coordinates in a vertical, branch-like manner. In normal communication (written and verbal), a chromosomal coordinate is likely to just be referred to as a band or simply the arm and number, e.g., "band p22.3" or just "p22.3." (chr = chromosome)

1. Heat denaturation into single strands
2. Binding of a primer sequence to its complementary target, usually upstream of the region of interest
3. Addition of DNA replication enzyme (DNA polymerase) to add nucleotides to the 3′ end of the primer into a growing chain of second strand DNA
4. Nucleotide analogues (**dideoxynucleotides**; **ddNTPs**) are incorporated into the replicating second strand of DNA; however, their analogue structure terminates the growing strand at their site of incorporation. The four nucleotides are differentially labeled (radioisotopes on separate lanes in the early years, replaced by fluorescent labels over the past two decades).
5. Electrophoresis (automated through capillary methods in the last two decades) to separate all the fragments according to size, allowing the sequence of nucleotide incorporation to be determined

Sanger sequencing is automated via capillary electrophoresis. However, its linear, serial nature means that long DNA sequences (Mb) can still require relatively long periods to be completed (weeks to months).

Pyrosequencing is a recent new platform for automated, rapid sequencing with fewer preparatory steps and quicker acquisition but generally shorter total lengths than Sanger sequencing. It measures differences in

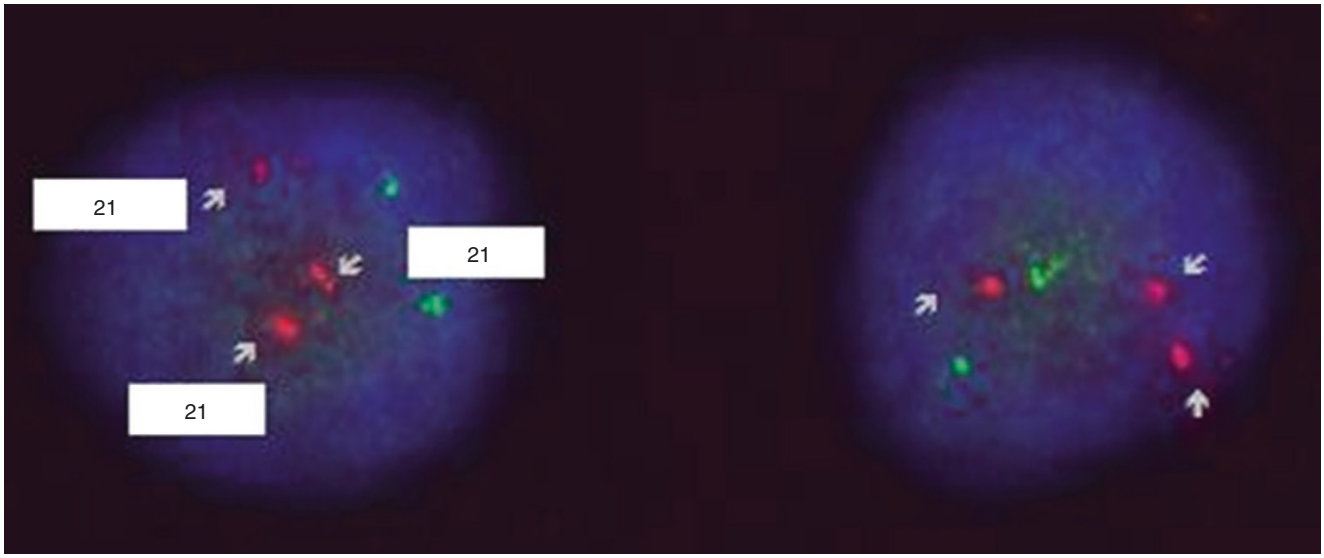


Fig. 3.10 Fluorescent *in situ* hybridization (FISH) of autosomes. Trisomy 21 indicated by *three red* fluorescently labeled copies of chromosome 21 (*arrows*) in two adjacent cells (Figure courtesy of Ms. R. Hutchinson, SA Pathology, Australia)

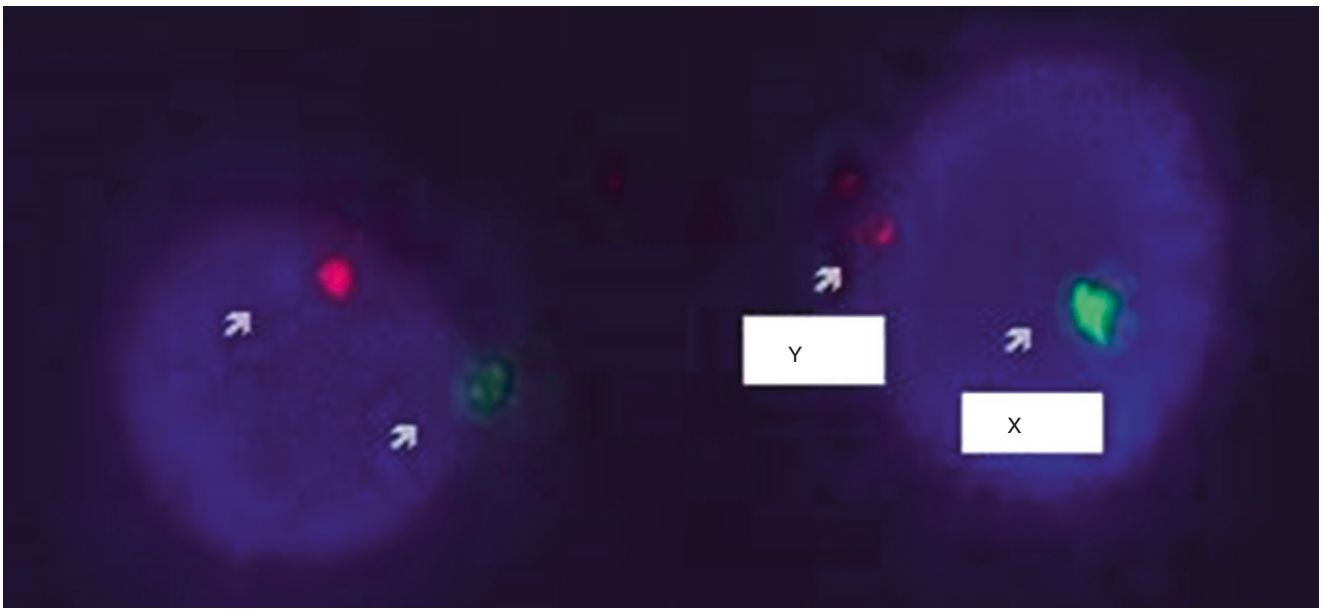


Fig. 3.11 Fluorescent *in situ* hybridization (FISH) of sex chromosomes. Male sex indicated by one copy each of the X (*green* fluorescence) and Y (*red* fluorescence) chromosomes (*arrows*) in two adjacent cells (Figure courtesy of Ms. R. Hutchinson, SA Pathology, Australia)

pyrophosphate release, between the four nucleotides as they are added to a replicating strand of DNA using chemiluminescence.

Output from automated sequencing is in the form of electrophoretic (Sanger; Fig. 3.12b) or light signal (pyrosequencing) spectra. It should be noted that this technology essentially produces a sequence that is an average (mean) of all the DNA molecules in the sample; therefore, changes that are only a small percentage of the whole (e.g., low-level mosaicism or somatic mutation) are difficult to detect by this method.

- **DNA sequencing** is a good test for the identification of many syndromes. Examples include Meckel-Gruber syndrome, osteogenesis imperfecta, and achondroplasia.

Restriction Fragment Analysis

This technique relies on cutting enzymes ("restriction enzymes") that cleave double-stranded DNA molecules at

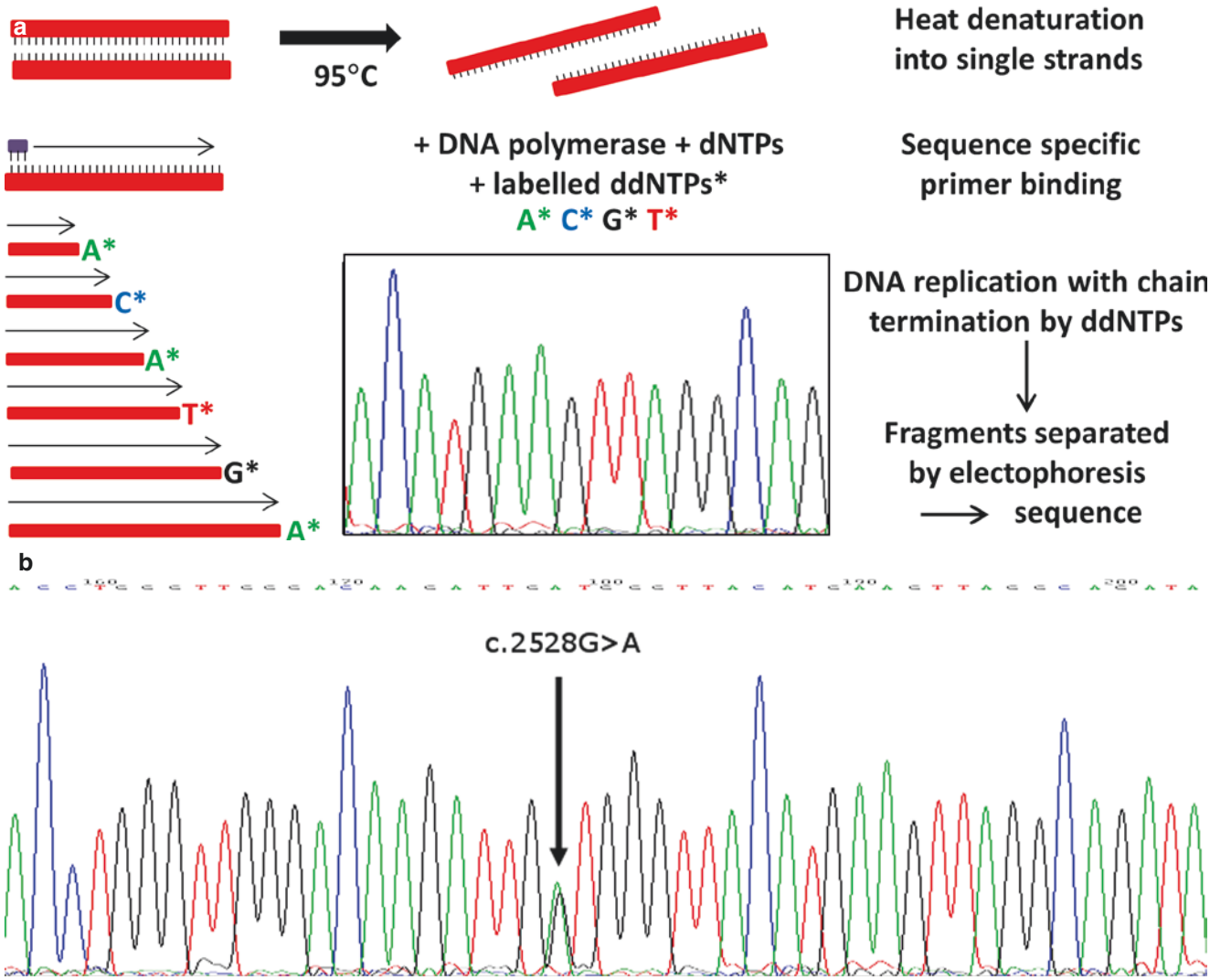


Fig. 3.12 DNA (Sanger) sequencing. (**a**) Method: Sanger sequencing utilizes labeled dideoxynucleotides (ddNTPs) to terminate chains of replicating DNA, initiated by a sequence specific primer (*purple*). This generates many DNA fragments that differ in size by only 1 bp, with their last incorporated nucleotide labeled. Separating these fragments according to size by electrophoresis allows a profile of the last incorporated nucleotide to be determined alongside the fragment just 1 bp shorter than it. A linear harvesting of DNA sequence data from a barcode-like readout of adjacent fragments is thus possible. This process is made much easier today by automation of DNA sequencing, furnished by fluorescently labeled ddNTPs, capillary electrophoresis, and software-based sequence analysis. (**b**) Example of a sequencing readout: DNA sequencing spectra (capillary electrophoresis) indicating a heterozygous mutation G>A (*black arrow*). At this position, there are two peaks of similar height—*green* (*A*) and *black* (*G*)—indicating the presence of both the normal sequence (*GGT*) on one allele and the pathogenic (mutated; GAT) sequence on the other allele (*PEX1* gene: c.2528G>A; p.G843D) (**b**) Courtesy of Mr. T. Pyragius, SA Pathology, Australia)

specific sequences. Recognition sites are usually short (4–8 bp; e.g., the enzyme EcoRI only cuts DNA at sites with the sequence GAATTC) and their frequency—i.e., number of times they cut—is often characteristic in a particular gene. If mutations occur in these recognition sites, they change the number of times the restriction enzyme cuts. Ultimately this leads to a difference in the number and size of fragments of DNA when separated by electrophoresis, giving a different banding pattern, called restriction fragment length polymorphism (RFLP). Amplification fragment length polymorphism (AFLP) relies on the generation of amplified polymerase chain reaction (PCR) products after restriction enzyme cutting of DNA, followed by ligation of specific PCR primers to the cut fragments. This enables only cut fragments to be subsequently amplified in a PCR reaction. The principle of generating a range of different sized fragments that characterize the presence or absence of a mutation is however overall the same as for RFLP.

The power of PCR (see later text) in tandem with restriction fragment analysis, in a technique called cleaved amplified polymorphic sequence (CAPS), is more commonly utilized today. Initially, PCR is used to generate a shorter fragment from a well-characterized region of interest using PCR. Restriction enzyme treatment then cuts the PCR product into separate smaller fragments according to the presence or absence of a mutation (Fig. 3.13).

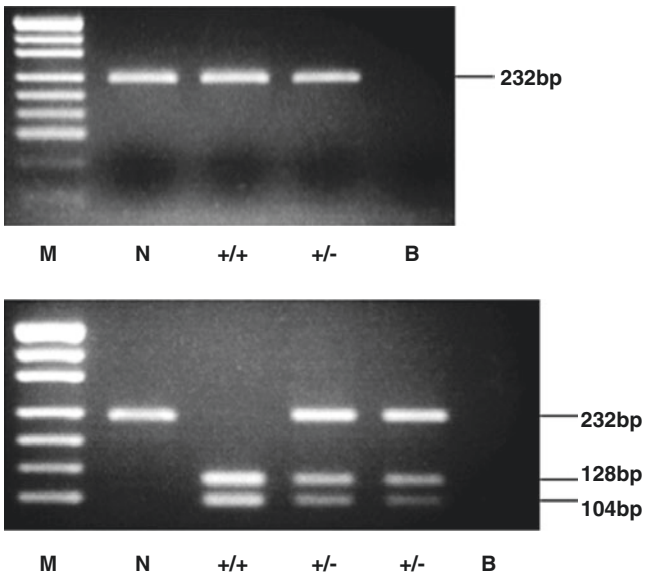


Fig. 3.13 Restriction fragment analysis of a PCR amplicon. An example of cleaved amplified polymorphic sequence (CAPS). PCR primers targeting a region of the *PMM2* gene amplify a 232 bp product in all samples, visualized on agarose gel electrophoresis (*upper panel*). Differences in DNA sequence produce differences in the ability for restriction enzymes to cut at their specific sequence targets. Differences in the DNA fragment profile after restriction enzyme digestion are referred to as CAPS. Shown in the *lower panel* is a restriction analysis-based method that detects a pathogenic mutation in the *PMM2* gene, associated with the condition congenital disorder of glycosylation type 1a (*CDG-1a*). Restriction enzyme BtsC1 cuts only at a single-nucleotide polymorphism in the amplified region of the *PMM2* gene. PCR products are cut into two smaller fragments only if this mutation is present. Individuals that are heterozygous for this mutation will have both the uncut (232 bp) and cut fragments present (128 bp and 104 bp). *M* molecular weight markers; *N* normal, no mutation (–/–); +/+ homozygous mutation; +/– heterozygous mutation; *B* blank (no DNA) control (Figure courtesy of Mr. K. Brion, SA Pathology, Australia)

- **RFLP/AFLP** is sometimes used to diagnose spinal muscular atrophy prenatally.

Linkage Analysis

The principle behind linkage analysis is explained in the section on gene structure (page 48). Essentially it uses alleles that are commonly inherited together as markers for specific genes, although they are unlikely to be the actual disease cause. These marker regions may be detected by DNA sequencing, RFLP, AFLP (see previous text), PCR, or Southern blotting (see later text).

In the book analogy, linkage analysis is equivalent to checking if an issue of a journal is present by checking whether other issues of the journal are collected together on the same shelf.

Historically, linkage analysis was responsible for the discovery of many genes (e.g., *CFTR*); however, the increasing availability of SNP arrays and exome- and genome-wide association studies using newer technologies will likely see the use of this technique continue to decrease.

Southern, Northern, and Western Blots

As described in the historical timeline (Fig. 3.1), this technique was named after its developer, Edwin Southern, not a map direction, hence the capitalization of "Southern." It was the first time that the techniques of complementary hybridization and fixation of DNA to a solid substrate after separation by electrophoresis were combined.

To use the book analogy, Southern blotting is like performing an online keyword search to highlight specific phrases or passages in book.

The same principle underlying this technique was then used for protein (Western blot) and RNA (Northern blot), a play on words from the map direction nuance. Like FISH (see previous text), all of these techniques rely on labeled probe hybridizing to a region of interest after electrophoresis and immobilization on a solid substrate (Fig. 3.14). Like FISH, Southern blotting and Northern blotting use a complementary nucleic acid, while Western blotting uses an antibody to the epitope of interest as the probe. The size of a nucleic acid probe and therefore the region of its complementary binding may be small (oligonucleotide) or very large (cDNA).

- **Southern blot** is commonly used to determine the length of a repeat sequence in fragile X syndrome or congenital myotonic dystrophy.
- **Western blot** is a good test for HIV antibody test confirmation.

Polymerase Chain Reaction (PCR)

Most genetic testing technologies used today rely on amplification of identical copies of a DNA region of interest from relatively small amounts of starting material.

Polymerase chain reaction (**PCR**) is the technology that underpins this amplification. Invented by Nobel laureate Kary Mullis in the mid-1980s, it essentially harnesses the inbuilt machinery of DNA replication, revolutionizing molecular biology to this day.

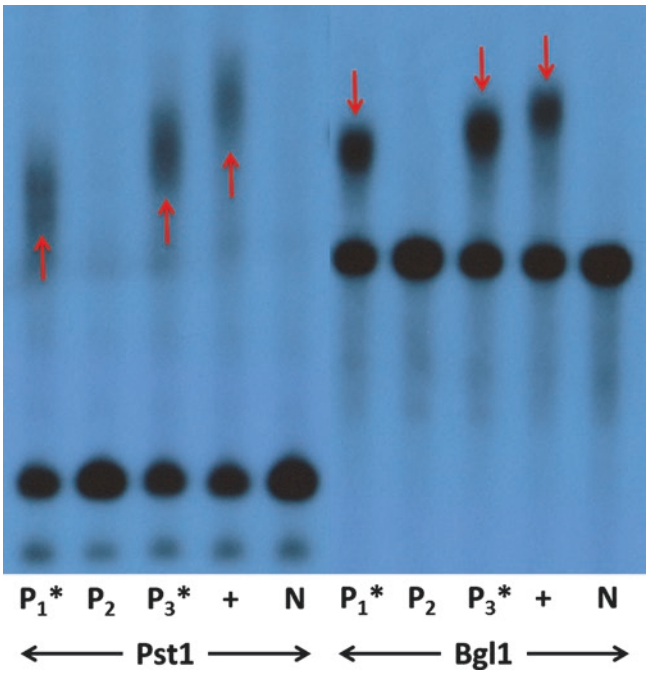


Fig. 3.14 Southern blot. DNA is cut into smaller fragments by restriction enzymes (here Pst1 and Bgl1) that only cut at specific recognition sequences, then electrophoresed on agarose gel, and transferred (blotted) onto a nitrocellulose sheet. A radiolabeled piece of DNA specific to the gene or region being probed hybridizes to regions containing complementary DNA (here M10M6 probe for the *DMPK* gene). The size of DNA fragments is estimated by how far they migrate from the origin during electrophoresis (larger fragments migrate more slowly, here closer to the *top*). *Red arrows* indicate restriction fragments from one allele that are greater in size than the normal range. Expansion of the number of CTG repeats in the noncoding region of the *DMPK* gene is associated with the autosomal dominant disorder myotonic dystrophy type 1 (DM1; normal 5–37, premutation 38–49, mild 50–150, classical 100–1,500, congenital 1,000–2,000 CTG repeats). Number of CTG repeats can be determined from the size of labeled DNA fragments. P_1 = approx. 1.2–2.2 kb fragment (412–743 CTG repeats), P_3 = approx. 1.9–2.7 kb fragment (629–904 CTG repeats), positive control = approx. 2.6–4.2 kb (867–1,400 CTG repeats). *P* patient sample, * pathogenic CTG expansion present, + positive control, *N* normal control (Figure courtesy of Ms. R. Catford and Dr K. Friend, SA Pathology, Australia)

Using the book analogy, PCR is like taking a specific paragraph or page from book, putting it on the photocopier, and setting the copy number to one billion.

Relying on variability in the strength of DNA binding to its complementary nucleotide sequence at different temperatures, PCR utilizes tightly controlled automated temperature cycling and a special heat-tolerant form of DNA polymerase (*Taq*—isolated from the thermophilic bacterium *Thermus aquaticus*) to rapidly and exponentially replicate specific sequences of DNA.

PCR consists of three phases repeated many times to exponentially amplify the target (Fig. 3.15a):

1. Denaturation of DNA into single strands at high temperature (>90 °C) to enable access for replication.
2. Binding of short specific single-stranded DNA sequences (primers) complementary to the region of interest at both 5′ and 3′ ends of the region to be amplified. This occurs at a temperature very close to the primer binding limits, ensuring a primer will bind only to its complementary sequence, giving it target specificity.
3. Elongation—*Taq* polymerase incorporates complementary nucleotides to the end of the growing chain extending from the primer to produce new double-stranded DNA molecules of the length bound by the primer pair.

The amplified product is referred to as an **amplicon**. Amplification of nucleic acids by PCR has many variations. Three of the most important variations (gap-PCR, long-range PCR, and MLPA) are discussed. However, direct differences in the size of PCR amplicons alone can be used to detect well-characterized genetic variants (Fig. 3.15b, c). Sequencing and MLPA (below) are often used as subsequent confirmatory methods following positive PCR results.

- **PCR-based amplification** is used, at some stage, in most genetic tests. It is often confused as "the" genetic test itself, but invariably its primary use is to amplify enough DNA to do "the" test.

Gap-PCR

This form of PCR relies on well-characterized deletions, bringing previously distant sequences very close together, so that primers to those sequences are close enough to now be successfully amplified by PCR.

- **Gap-PCR** is a good test for detecting hemoglobinopathies, such as Hb Barts in hydrops fetalis (Fig. 3.16a).

Long-Range PCR (LR-PCR)

In standard PCR, there is an underlying error rate for misincorporation of nucleotides (of the order of once per 10,000–100,000 nucleotides). *Taq* polymerase stalls to correct these errors. The longer a DNA strand, the more likely there will be errors and the efficiency of replication compromised by *Taq* stalling to repair them. This sets a practical limit to the length of DNA able to amplified using standard PCR to a few thousand base pairs.

Incorporation of a proofreading enzyme into a PCR mix helps to iron out these errors earlier, allowing *Taq* and/or other DNA polymerases to produce longer amplified products of the order of tens of kilobases. This is called **long-range PCR** (**LR-PCR**). It is used in applications where amplification

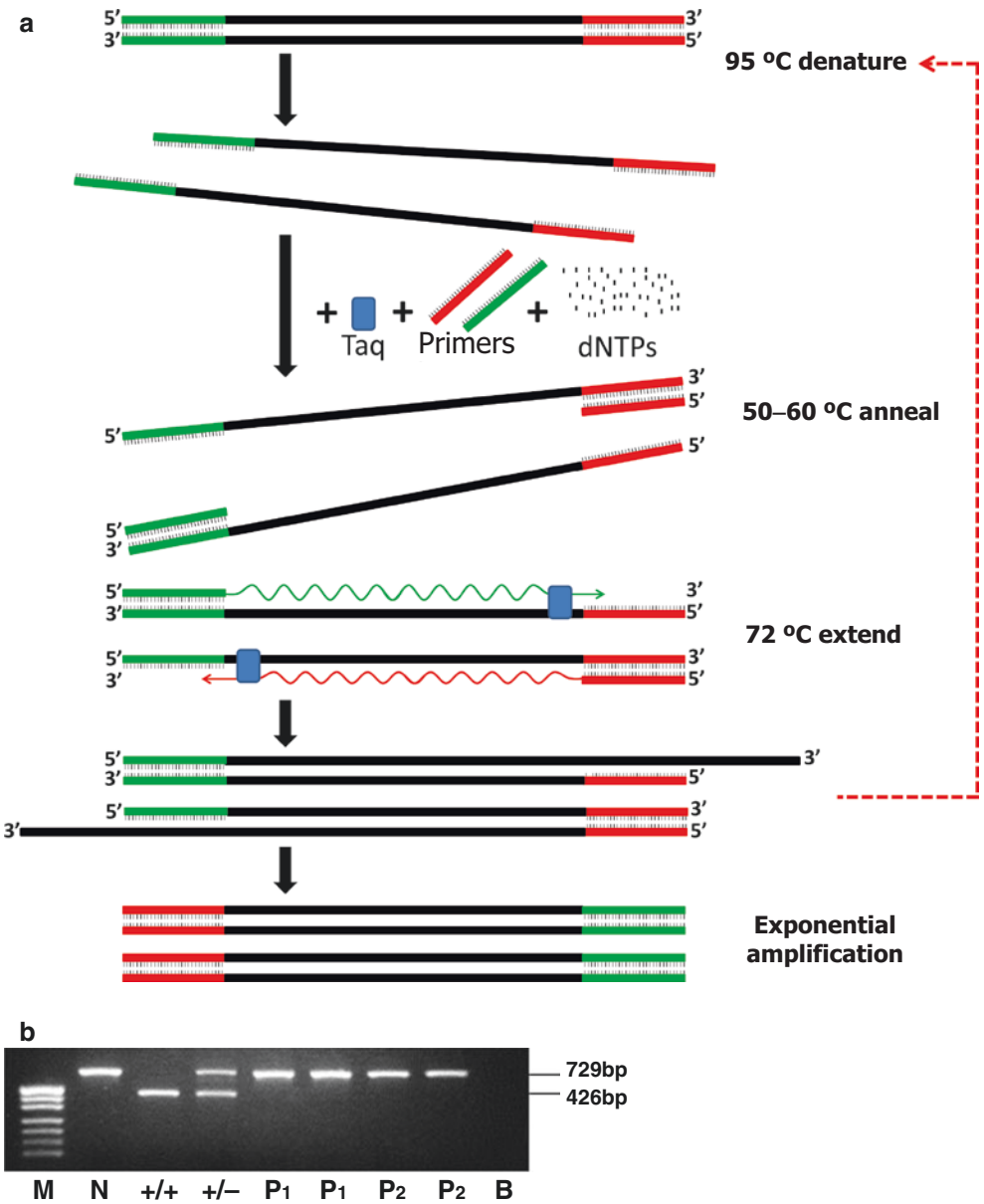


Fig. 3.15 Polymerase chain reaction (PCR). (**a**) PCR amplification of DNA. DNA replication requires a DNA polymerase, primers to initiate the region of replication, and nucleotides (dNTPs, the nucleotide building blocks of DNA). In PCR, the steps of denaturation, annealing of primers, and extension of sequence from the primers happen at tightly controlled temperatures. *Thermus aquaticus* (*Taq*) polymerase and other DNA polymerases that can perform and survive at relatively high temperature allow rapid cycling of these steps to produce an exponential amplification of target DNA. Many modern techniques of genetic testing are entirely reliant on a DNA amplification step, underpinned by PCR-like procedures. (**b**) PCR analysis of a gene deletion (agarose gel electrophoresis). Differences in the length of DNA of a PCR-amplified product can indicate deletions or duplications to that region. Differences in the profile of PCR-amplified products are visualized by electrophoresis, separating amplicons according to size. Shown here on agarose gel electrophoresis is a 203 bp decrease in the size of the PCR-amplified product targeting a pathogenic deletion in the *CLN3* gene (associated with ceroid lipofuscinosis, neuronal, type 3 [Batten disease]). Individuals that are heterozygous for this mutation will produce PCR-amplified products both with (426 bp) and without (729 bp) the deletion. *M* molecular weight markers; *N* normal control (no mutation, –/–; 729 bp); +/+ homozygous mutation (426 bp); +/– heterozygous mutation (426 and 729 bp); *P1,2* two normal patient samples (no mutation, –/–, 729 bp); *B* blank (no DNA) control (**b** Courtesy of Mr. K. Brion, SA Pathology, Australia). (**c**) PCR analysis of CGG repeats in fragile X syndrome (capillary electrophoresis). PCR primers target the CGG repeat region of the *FMR1* gene on the X chromosome, associated with fragile X syndrome (FXS). Capillary electrophoresis differentiates PCR products according to size, allowing the number of CGG repeats to be determined (normal 5–44, gray zone 45–54, premutation 55–200, FXS >200). Shown here are 30 CGG repeats in a male (only one X chromosome; *upper panel*), an unaffected female (22 and 29 repeats; *middle panel*), and a female normal on one allele and a premutation on the other allele (29 and 54 repeats; *lower panel*). This method will not detect deletion or missense mutation causes for FXS (**c** Courtesy of Dr. K. Friend, SA Pathology, Australia)

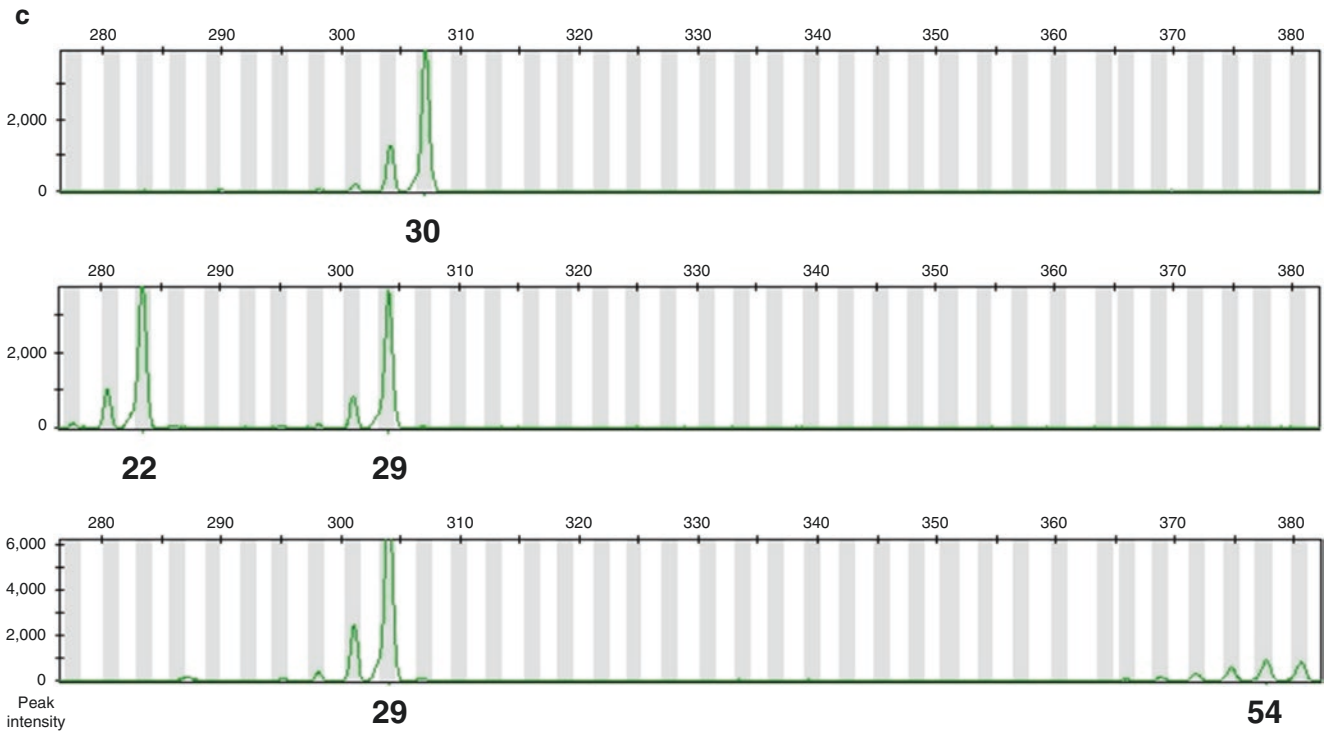


Fig. 3.15 (continued)

with good fidelity over larger stretches of DNA is required, e.g., complete mitochondrial DNA sequence.

- **LR-PCR** is a good test for detecting incontinentia pigmenti (Fig. 3.16b).

Multiplex Ligation-Dependent Probe Amplification (MLPA)

A PCR technique, **MLPA** is used to detect copy number variations (deletions or duplications) in genes (see discussion in array, page 73). It uses one primer pair to amplify PCR products from multiple regions in one reaction. Each region produces a uniquely sized amplicon due to differences in the size of stuffer and gene-specific regions of hybridization probes but flanked by the same common primer sequence that will amplify in PCR amplification. If an exon or part of it is missing, then that region will not be amplified in an MLPA reaction. The specificity of this technique lies in the fact that amplification will only be successful if sequence is identified where the probes sit adjacent to each other, so that a small gap between them can be filled in by the enzyme ligase that then allows PCR amplification to proceed (Fig. 3.17a).

A control sample known to amplify all the regions being assessed is used as a reference, in a similar manner to CGH array (page 73), to detect copy number variations (CNVs).

Good examples of its application are given in references [40–42].

To use the book analogy, MLPA is like seeking out two adjacent words in a book. You only get a hit if the word order and combination is an identical match. Output is similar to getting frequency count for the word combination, broken down according to chapter or subsection.

- **MLPA** is a good test for Duchenne muscular dystrophy (DMD), fragile X syndrome, and microdeletion syndromes, e.g., DiGeorge (22q11.2) syndrome and spinal muscular atrophy (SMA) (Fig. 3.17b–d).

Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) Mass Spectrometry

This technology has been adapted for use in identifying many biomolecules. The most common use in genetic testing is looking for well-characterized single-nucleotide mutations in DNA. It begins with a PCR amplification step to generate starting material specific for the gene of interest. In a second separate reaction, there is extension of one single nucleotide onto

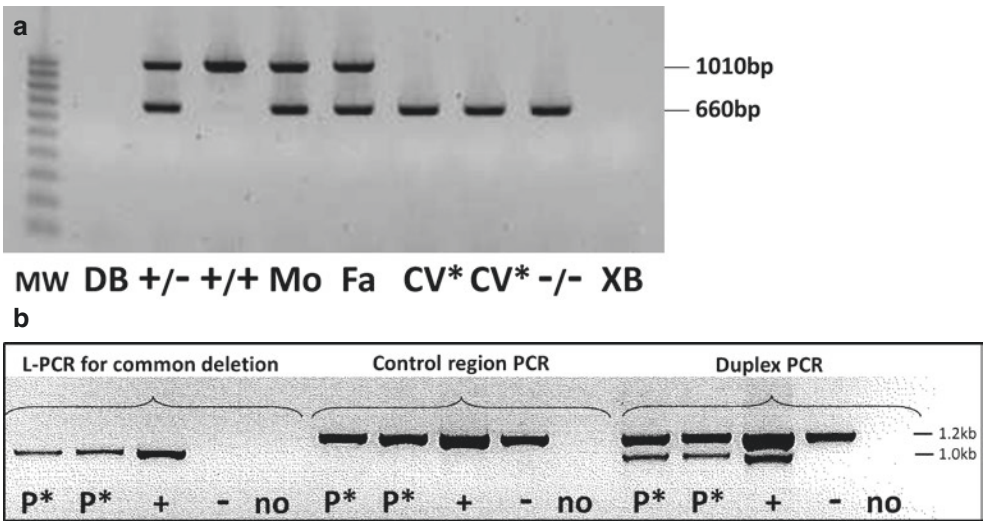
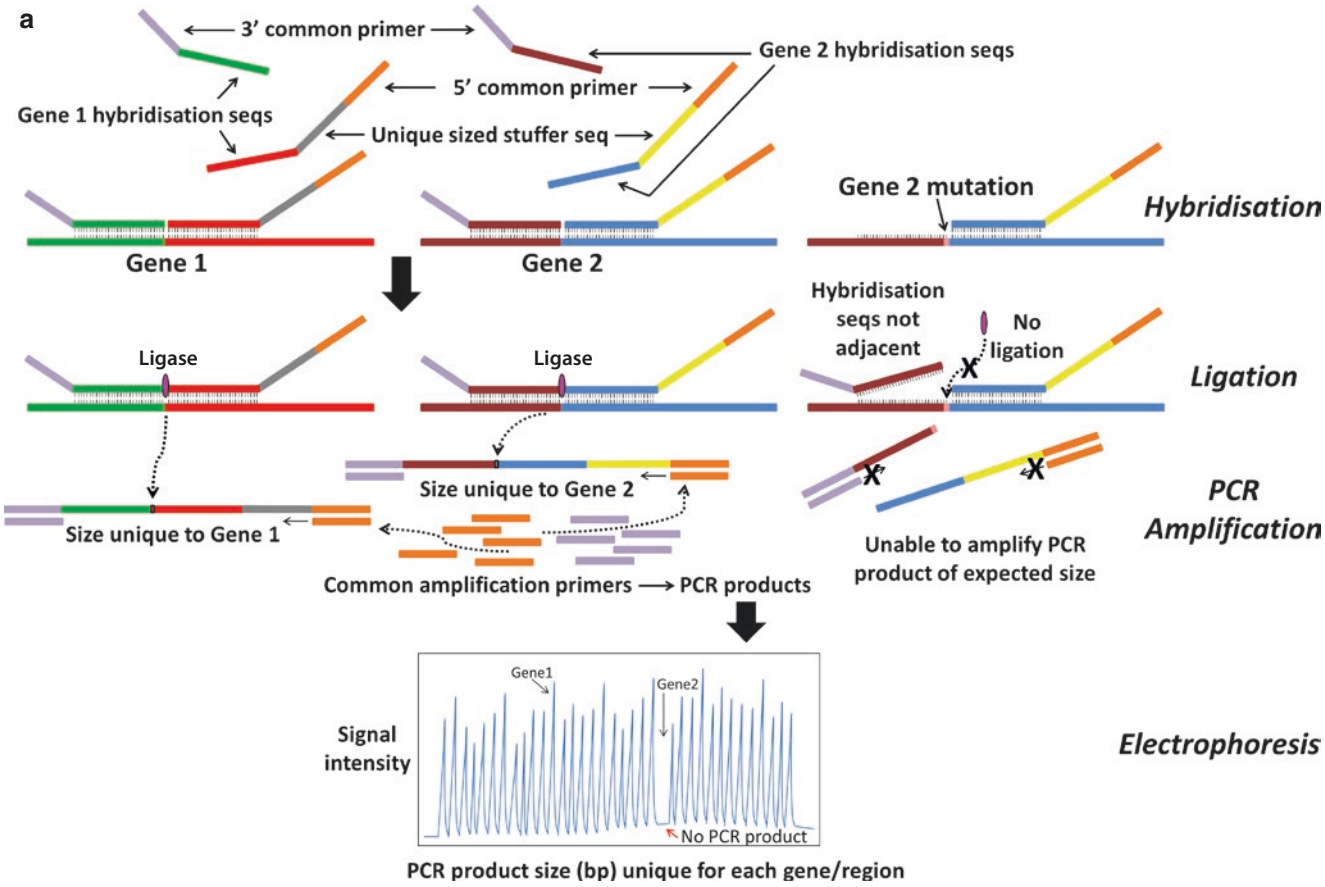
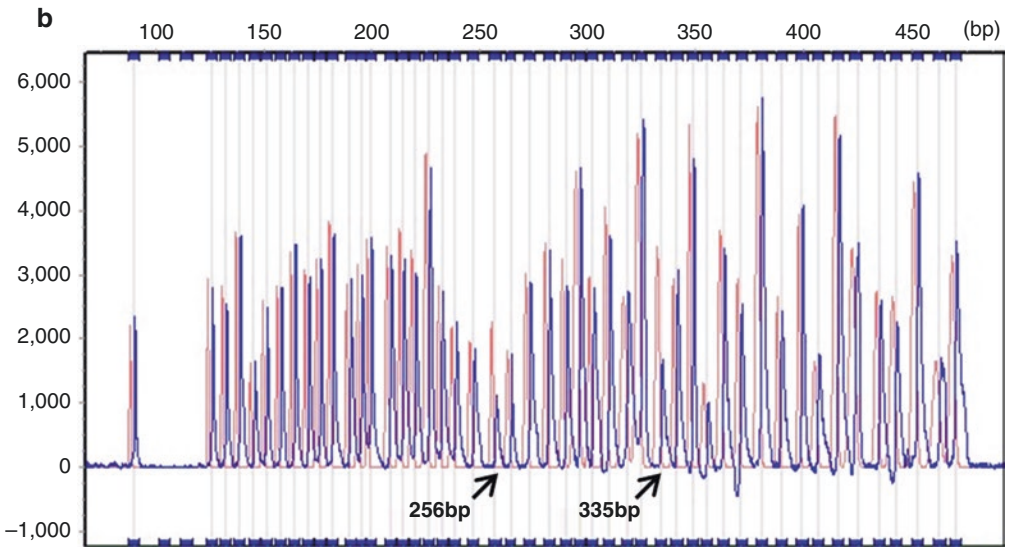


Fig. 3.16 (**a**) Gap-PCR analysis of alpha thalassemia. Two closely located genes encode alpha globin (*HBA1* and *HBA2*; both on chr16p13.3). Common deletion mutations in alpha globin can be detected by gap-PCR. Pathogenic deletions in these genes result in various forms of alpha thalassemia, depending on the number of functional alpha globin alleles (normal = 4, one from each gene on each allele). Homozygous deletion mutations on both alleles for both genes result in no functional alpha globin protein (Hb Barts, causing fetal demise from hydrops fetalis). Shown here is one gene deletion found predominantly in those of Southeast Asian ethnicity that spans both alpha globin-encoding genes. This deletion involves the removal of about 19.4 Kb of DNA. Gap-PCR produces a smaller amplicon if a deletion mutation is present—wild type (no deletion) 1,010 bp, heterozygous deletion mutation both 1,010 bp and 660 bp, homozygous deletion mutation only smaller 660 bp amplicon. Both parents are seen to be heterozygous for this mutation, with their fetus being affected (homozygous; Hb Barts). *MW* molecular weight markers; *DB* DNA blank control for PCR; +/– heterozygous deletion mutation control; +/+ wild-type (no deletions) control; *Mo* mother; *Fa* father; *CV** fetal chorionic villous sample; –/– homozygous mutation control; *XB* DNA extraction blank control (Figure courtesy of Dr. K. Simons and Dr C. Nicholls, SA Pathology, Australia). (**b**) Long-range PCR (LR-PCR). Conventional PCR utilizes thermostable *Taq* polymerase for amplification of DNA targets. *Taq* allows rapid amplification but has limitations on the maximum size of the amplified product. Long-range PCR utilizes high-fidelity DNA polymerases with proofreading ability to allow amplification of very long DNA fragments (up to 40 kb). The most common deletion in the *IKBKG* gene (associated with incontinentia pigmenti) is 11.7 kb (spanning exons 4–10). The markedly decreased size of the amplicon containing this deletion (1.0 kb) is shown here from an LR-PCR reaction (*left*). Samples without the deletion will produce a much larger amplification product (~13 kb; not shown) and no smaller 1.0 kb amplification product. An unrelated, ubiquitous region of DNA acts as an amplification reaction control (1.2kb; *middle*). Duplex PCR combines both the *IKBKG* gene and control primers in the same tube to control for any differences in amplification efficiency (*right*). *P** patient with mutation present, + positive control, – negative control, *no* no DNA control (Figure courtesy of Dr. K. Friend, SA Pathology, Australia)

Fig. 3.17 (**a**) Multiplex ligation-dependent probe amplification (MLPA). Many mutations in multiple genes (or even different regions of the same gene) can be tested in the same single reaction tube (multiplexed). Common PCR primer sequences (*violet* and *orange*) flank hybridization sequences (seqs; Gene 1, *green* and *red*; Gene 2, *brown* and *blue*) specific for individual gene mutations. A ligation step after hybridization will only occur if both hybridization sequences (probes) for that region completely hybridize, so that they lie adjacent to each other. DNA ligase (*pink*) is then able to fill in the gap between these adjacent hybridized probes (*black rectangle*). This allows the common primers to amplify a PCR product from any region that has completely hybridized to their hybridization probes. Any mutation (Gene 2, *light pink*; *far right*) will not allow complete hybridization of the hybridization probes resulting in failure of the ligation step and subsequent failure to amplify a PCR product for that region. The combination of stuffer sequence (Gene 1, *gray*; Gene 2, *yellow*) and hybridization sequence is designed so that each gene region produces a uniquely sized PCR amplification product, when analyzed on capillary electrophoresis (*bottom*). In this way, many genes (or regions from the same gene) can be analyzed simultaneously in the one reaction. Similar to CGH array (Fig. 3.21), comparison of copy number variation (*CNV*) between a control and test sample analyzed by MLPA indicates if there have been deletions or duplications of DNA but over much smaller regions (50–70 bp) than possible with CGH array. (**b**, **c**) MLPA analysis of microdeletion syndromes. Analysis of 20 microdeletion syndromes simultaneously, using a commercial MLPA kit (P0245; MRC Holland). PCR-amplified products are separated by size on capillary electrophoresis. Amplified product size and relative quantity from a test sample (*blue trace*) is compared to a normal control (*red trace*) to determine copy number variations (CNVs) in amplified regions. (**b**) Heterozygous deletions of two probes (256 and 335 bp) within the region associated with autosomal dominant neurofibromatosis type 1 (*NF1*) microdeletion syndrome are indicated by a reduction of *blue* trace to approximately half of the *red* trace peak height for the size of amplified product expected for this region (*arrows*). (**c**) The same data can be presented as a peak ratio to more clearly delineate CNVs. Peak ratios of approximately 1 indicate no CNV (*green boxes*). Peak ratios greater than 1.25 or less than 0.75 (*green horizontal lines*) suggest marked CNV, i.e., duplications or deletions, respectively. Heterozygous deletion is indicated by a peak ratio of approximately 0.5, corresponding to an expected decrease in amplified product by one half if it has been deleted from one of a pair of alleles. The probes deleted are at chr17q11.2 within exons 12 and 20 of the *NF1* gene (*red boxes*, *arrowed*). (**d**) MLPA analysis for spinal muscular atrophy. MLPA peak ratio analysis indicates homozygous deletion of regions of the *SMN1* gene (peak ratio = zero, indicating no amplified product detected in this region, i.e., deletion from both alleles). Deletions of two probes to exons 7 and 8 (182 and 218 bp, respectively; *red boxes*, *arrowed*) of the *SMN1* gene are associated with the autosomal recessive condition spinal muscular atrophy (SMA) (**b**–**d** Courtesy of Dr. K. Friend, SA Pathology, Australia)


a
3′ common primer
Gene 2 hybridisation seqs
5′ common primer
Gene 1 hybridisation seqs
Unique sized stuffer seq
Gene 1
Gene 2
Gene 2 mutation
Hybridisation
Ligase
Ligase
Hybridisation seqs not adjacent
No ligation
Ligation
Size unique to Gene 2
Size unique to Gene 1
Unable to amplify PCR product of expected size
PCR Amplification
Common amplification primers → PCR products
Gene1
Gene2
Signal intensity
No PCR product
Electrophoresis
PCR product size (bp) unique for each gene/region


b
100
150
200
250
300
350
400
450
(bp)
6,000
5,000
4,000
3,000
2,000
1,000
0
−1,000
256bp
335bp

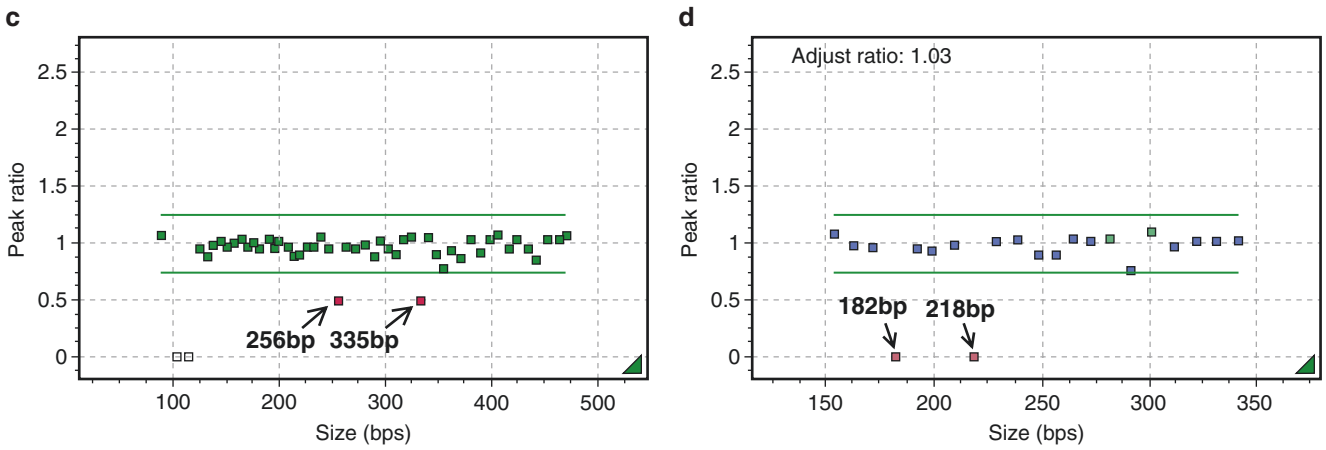


Fig. 3.17 (continued)

the amplified product using nucleotides modified to have a specific mass (Fig. 3.18a). Resulting samples are purified then spotted by a robotic device in nanoliter quantities onto a silica chip, much like in microarrays. This gives the advantage of very-high-density throughput so that many samples can be assessed in tandem. Firing of a finely controlled laser precisely onto each individual spot rapidly and sequentially converts it into ionized plasma for passing through a connected mass spectrometer, creating a mass particle profile for each sample. Mutations and normal sequence have characteristic mass particle signatures, assessed and called automatically in software.

A great advantage of this technique is that several different mutations can be assessed in the one tube, as long as each amplified, mass-labeled product has a unique mass compared to other products in the same tube. This multiplexing of both an increased number of individual samples in the one run and the number of mutations that can be assessed simultaneously has markedly increased the power and decreased the cost of this technology for mutation screening in conditions with high carrier prevalence (e.g., cystic fibrosis; Fig. 3.18b).

- **MALDI-TOF** is a good test for many of the common mutations found in the cystic fibrosis (*CFTR*) gene.

Mini-/Microsatellite Repeats

Satellite repeats are short sequences of DNA repeated next to each other (called **variable number tandem repeats**, **VNTRs**) at specific sites throughout the genome, often in noncoding regions. Microsatellites are repeats of 2–6 bp (**short tandem repeats**; **STR**), while minisatellites are longer VNTRs of 10–60 bp. The number of times the sequence is repeated in tandem is highly variable between individuals. PCR-based techniques can be used to amplify these repeat regions, and the number of times a VNTR is repeated can be determined from their size on electrophoresis. The number of repeats in several VNTRs will be characteristic for each individual and forms the basis of **DNA fingerprinting**.

As half of our genome is inherited from each parent, we also get half of our satellite repeat patterns from each parent. Therefore, this technique is useful for parentage analysis, e.g., in paternity cases or in determining the level of maternal cell contamination in a fetal sample (Fig. 3.19).

Satellite repeat results are generally presented as electrophoresis spectra indicating the number of repeats found in a range of different VNTRs in the same individual.

The technique is useful in molar pregnancy testing where misexpression of imprinted genes leads to a complete or partial hydatidiform mole. Determining parental origin of the imprinted genes is helpful for proper classification, determining likely pathology and most appropriate management [43].

- **Satellite repeat marker analysis** is a good test for maternal cell contamination, molar pregnancy, forensic identification, and paternity testing.

CpG Methylation

Data on CpG sites where cytosine is methylated to produce 5-methyl cytosine is used to determine regions of epigenetic gene silencing. Conversely, CpG hypomethylation at known MVPs indicates increases in gene expression at these sites (see epigenetics on page 57).

CpG methylation tests all rely on the treatment of genomic DNA with bisulfite (alkylation). Bisulfite converts cytosine to uracil but does not change 5-methyl cytosine (Fig. 3.20a).

High-resolution melting analysis, methylation-specific PCR, and standard PCR followed by MALDI-TOF, RFLP, or sequencing and methylation arrays are all techniques used to determine methylated CpG sites that are resistant to bisulfite treatment. The technique chosen depends on the length of coverage required.

- **CpG methylation analysis** is useful in assessing diseases related to genomic imprinting, such as Angelman and Prader-Willi syndromes (Fig. 3.20b) (see epigenetics on page 57).

Cytogenetic Microarray (CGH and SNP Array)

Microarrays are an important and natural choice as part of the fetal diagnostic process. They can indicate differences in chromosome structure at a higher resolution (0.1–1 Mb) than attainable by karyotyping. The American College of Medical Genetics (ACMG) 2010 review of clinical use of array-based technologies recommends them as a first-tier test for investigating developmental delay/intellectual disability, multiple congenital abnormalities, and autism spectrum disorders [44]. They cite evidence from large cohort studies estimating between 10 and 20 % improved diagnostic yield compared to karyotyping.

This technology relies on robotic workstations to spot well-characterized DNA fragments at very high density in specific order onto silicon microchips (**microarrays**). The entire genome of an individual, fragmented into smaller pieces, can then be applied to the chip where it will hybridize to its complementary sequence at a specific location, already mapped on the chip.

Differences in hybridization patterns between a test and reference genome indicate **copy number variation** (**CNV**), i.e., differences in the number of times one of the smaller fragments of DNA is present within the genome.

To labor the book analogy, arrays are like a whole library stocktake, attempting to identify any books missing or available in multiple copies.

The hybridization component of both **CGH** (**comparative genomic hybridization**) and **SNP** (**single-nucleotide polymorphism**) microarray techniques is analogous to 100,000s of FISH hybridization reactions being run in

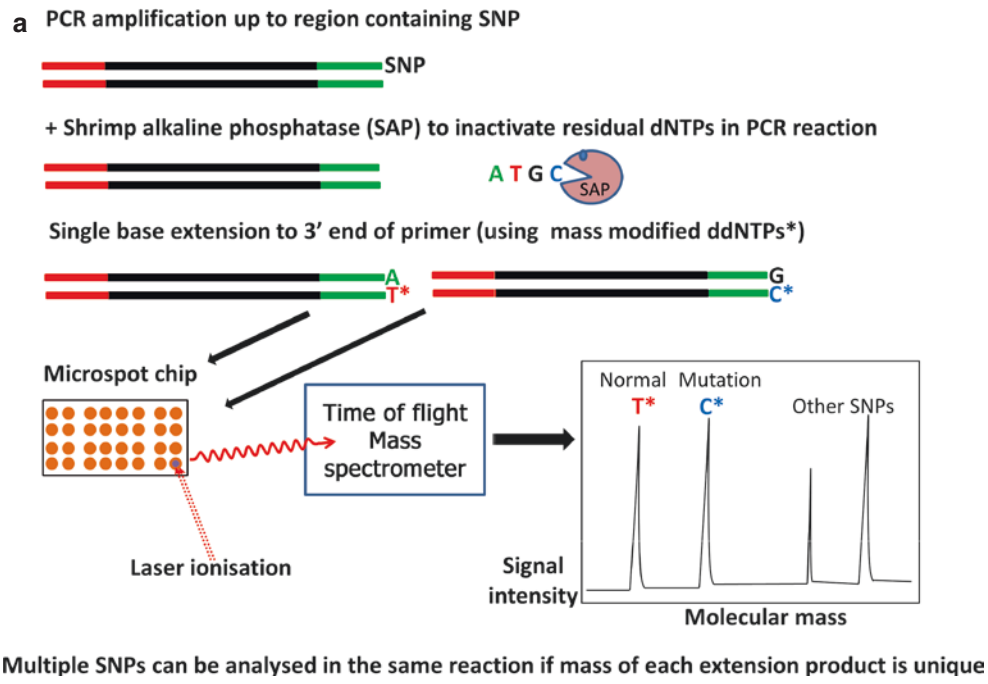
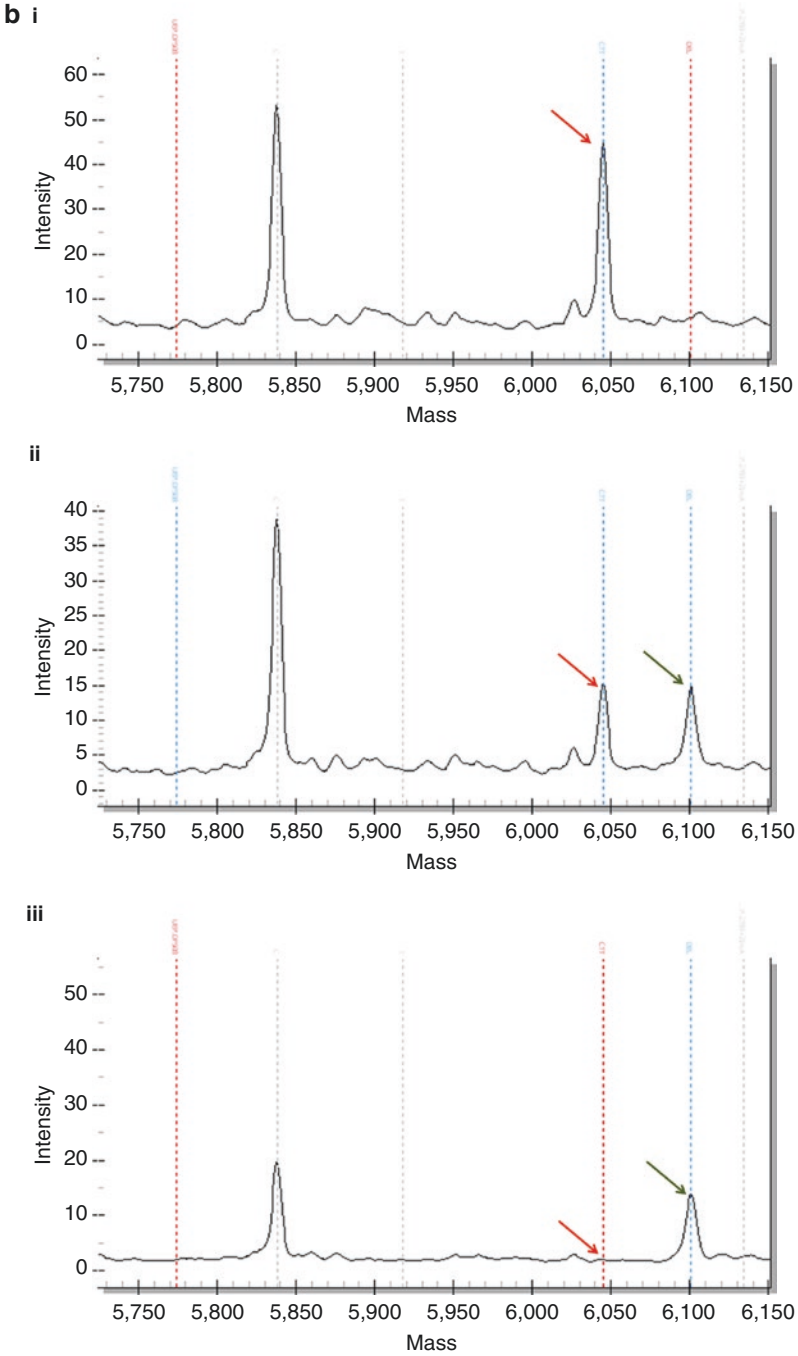


Fig. 3.18 Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry. (**a**) Method: this test relies on the incorporation of mass-modified ddNTPs to produce a specific mass spectrometric signature. It allows multiple single-nucleotide polymorphisms (SNPs) to be assessed in the same reaction. PCR generates an amplified product next to the SNP of interest. Cleanup of the PCR reaction with shrimp alkaline phosphatase (SAP) removes any remaining dNTPs so that they will not interfere with the subsequent single base extension step. The use of chain terminating ddNTPs with a modified mass ensures that only one single base extension will occur and that the extension product will have a unique mass based on the nucleotide incorporated. Microspotting onto a silicon chip, followed by laser ionization feeding directly into a mass spectrometer, allows rapid, automated analysis both of many SNPs in the one reaction and multiple samples spotted at high density onto the same microchip. (**b**) *CFTR* gene mutation c.1521_1523delCTT (p.F508del) in cystic fibrosis: (i) Normal (CTT intact; *red arrow*). (ii) Heterozygous (both CTT and deletion (DEL) with similar peak heights; *red* and *green arrows*, respectively). (iii) Homozygous (deletion (DEL) peak only with no CTT peak; *red* and *green arrows*, respectively) (**b** Courtesy of Mr. T. Pyragius, SA Pathology, Australia)

Fig. 3.18 (continued)



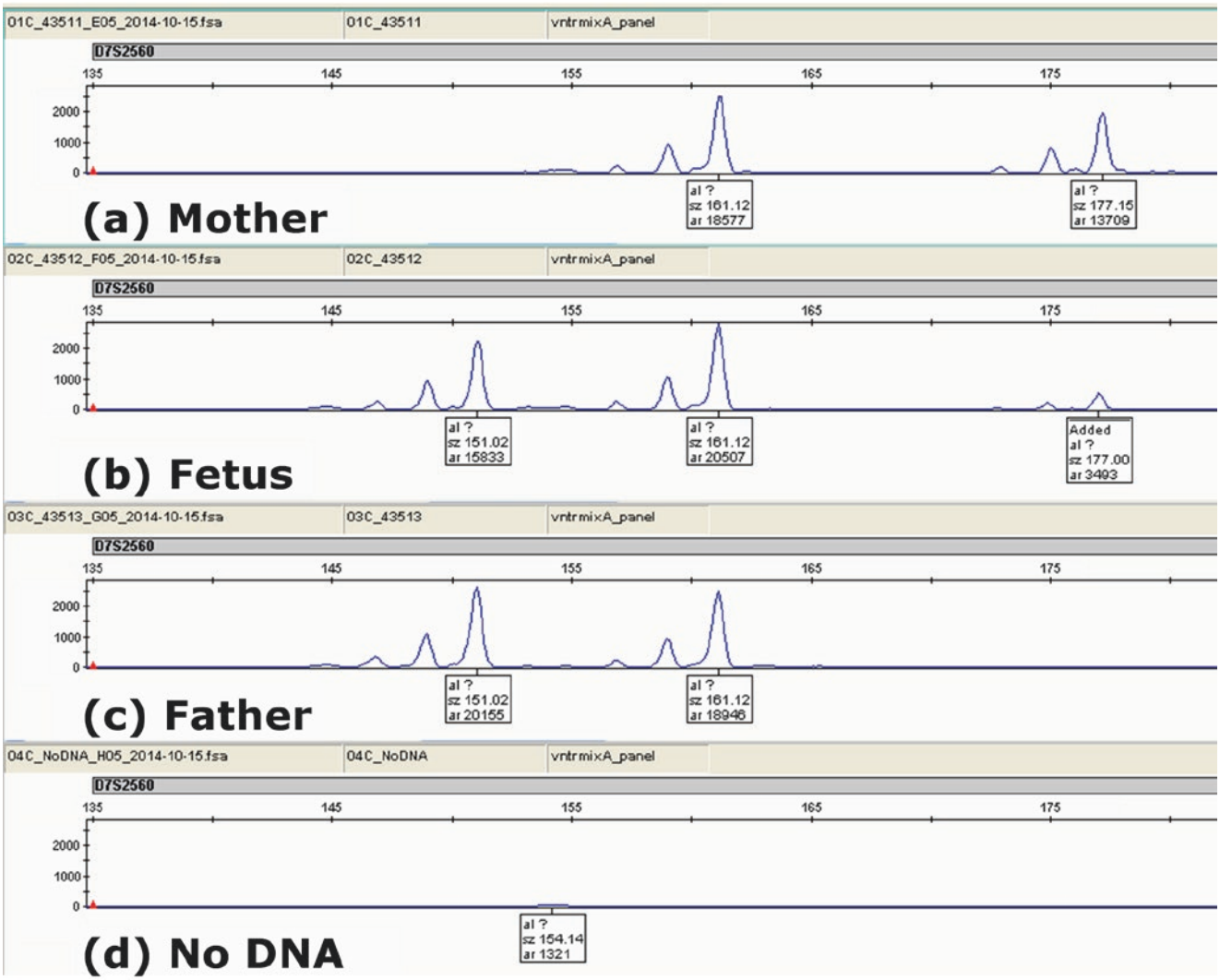


Fig. 3.19 Satellite repeat-based DNA fingerprinting for detecting maternal cell contamination. Sets of DNA microsatellite markers are amplified for maternal, paternal, and prenatal (fetal) samples. The distance between microsatellite markers and therefore size of amplified products will differ between individuals, acting as a unique DNA fingerprint. Peaks are separated according to molecular size. Although for this set of markers the mother and father share a common 161 bp marker on one allele, the father (**c**) has a 151 bp marker and the mother (**a**) a 177 bp marker on the other allele. The fetus (**b**) should only inherit one allele from each parent; however, there are three peaks present (151, 161, and 177 bp) indicating that the sample is contaminated with some maternal tissue. A no DNA control (**d**) does not produce any amplified products. Unrelated individuals may share some common microsatellite markers on one allele; however, using many sets of markers ensures a unique profile for each individual. A similar strategy is used for forensic DNA fingerprinting and determining parentage (Figure courtesy of Ms. R. Catford and Dr K. Friend, SA Pathology, Australia)

parallel next to each other on the one chip. Automated microscopic imaging and analysis is then used to determine fluorescent intensity at each spot, to assess differences in hybridization compared to a reference ("normal") genome.

The ability of CGH array to interrogate the whole genome was a spin-off of the methods used in early stages of the Human Genome Project, using bacterial artificial chromosomes (BACs) for cloning and sequencing. This generated an industrial-scale production of many smaller fragments covering the entire genome that were eventually utilized on microarrays. CGH relies on a test genome being fluorescently labeled a different color (green) to the reference genome (red). The two samples are then combined and hybridized to the microarray chip together. Identical sequences will hybridize to the same locations on the chip. Differences in signal intensity between different spots on the chip are easily evident when imaged; i.e., equal signal intensities will result in a yellow spot (combination of red and green). Spots that are more green or red indicate copy number variations between the test and reference genome (Fig. 3.21a).

SNP microarray, in contrast, utilizes hundreds of thousands of specific oligonucleotides, generated to cover the entire genome, with inclusion of many containing single-nucleotide polymorphisms (SNPs) known to be commonly

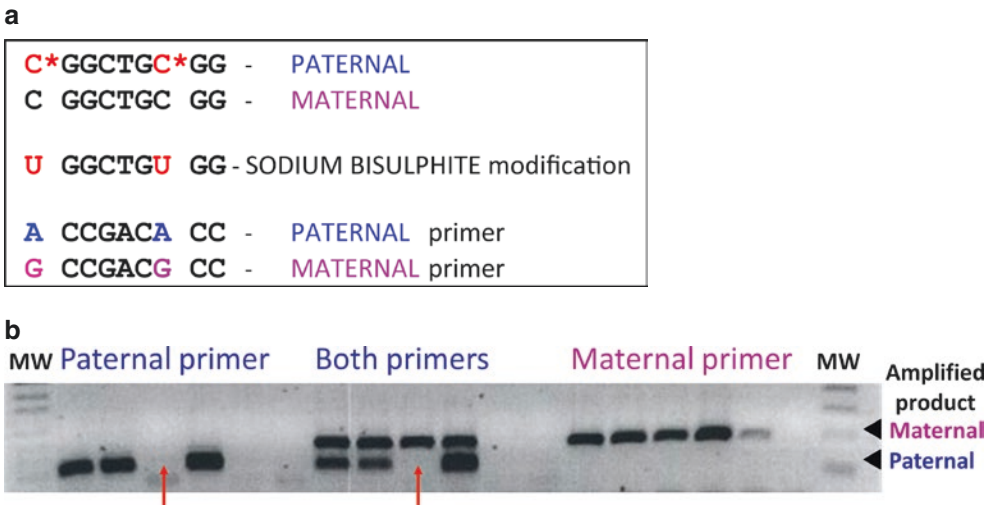


Fig. 3.20 Methylation PCR. (**a**) Paternal and maternal alleles have different methylation patterns (imprinting). Bisulfite alkylation of cytosine to uracil (U) does not occur at CpG methylated sites (C*). This enables design of primers that will only bind to nonmethylated regions after bisulfite alkylation. (**b**) This technique can be employed to determine imprinting patterns important in conditions such as Angelman and Prader-Willi syndromes (PWS), as well as other epigenetic modifications. Following bisulfite alkylation, PCR is conducted with primers specific for maternal non-alkylation and paternal alkylation products (**a**). Agarose gel electrophoresis indicates the absence of the paternal amplification product (*red arrows*) and the presence of the maternal amplification product, consistent with the paternal imprinting pattern found in PWS. *MW* molecular weight markers (Figures courtesy of Dr. K. Friend, SA Pathology, Australia)

associated with disease. This can increase the resolution and precision of the sequences found to have CNVs. SNP array does not use a reference genome in the same hybridization reaction; rather it compares the test genome fluorescent hybridization signal to an archived, well-characterized reference genome through software (Fig. 3.22a). Current commercially available SNP arrays use 850,000 unique oligonucleotides on their microchips.

Virtual karyotypes can be constructed in software from both CGH and SNP arrays (Figs. 3.21b, c and 3.22a) as the chromosomal location of probes used is well characterized, with coverage across all chromosomes.

Identified CNV regions may contain multiple candidate genes that could be causative for disease. Online tools such as the UCSC Genome Browser (Fig. 3.22b) and OMIM (see under nomenclature on page 54), plus biomedical literature searches, are used to help interrogate array results. Clinical phenotype needs to correlate with known or predicted pathogenic regions for results to be meaningful. Therefore, sometimes CNVs that are a true pathogenic region may be indicated from array results but not reported due to the lack of supporting evidence.

Loss of heterozygosity (**LOH**) refers to the deletion of an entire gene and/or surrounding chromosomal region, so that an allele from one parent is entirely lost. **Hemizygous** and **haploinsufficient** are related terms, used to describe a similar concept (see inheritance on page 55). Regions with LOH are worth closer examination as potential hotspots for disease, often through gene dosage effects, i.e., reduction in relative expression of a gene product, indicated by CNV.

CGH arrays are unable to determine balanced chromosomal anomalies (e.g., translocation, inversion, ring chromosome) or low levels of mosaicism.

The increased specificity of SNP compared to CGH arrays allows **copy neutral LOH** to be detected. Also known as **uniparental disomy** (**UPD**), it refers to replication of the same chromosome from one parent, after loss of the chromosome from the other parent, during early development. It is important as a potential hotspot for recessive allele expression (given both alleles are copies of each other and therefore automatically homozygous).

Array technologies are batched to reduce costs, with analysis taking variable periods dependent on the complexity of results. Therefore, turnaround time can be in the order of weeks to months, depending on the level of priority, but often useful when clinical condition warrants it and standard karyotyping has not detected any abnormalities. This technology can be adapted to assess more specific subsets of disease (e.g., developmental disorder arrays), and the technique is already in use for the assessment of CpG methylation.

- **CGH arrays** are a good test to identify the cause of congenital abnormalities and intellectual disability (10–15 % more chromosomal diagnoses made if the standard karyotype is normal).
- **SNP arrays** are also a good test for identifying the cause of congenital abnormalities but have the added benefit of providing extra information about potential recessive diseases in consanguineous couples.

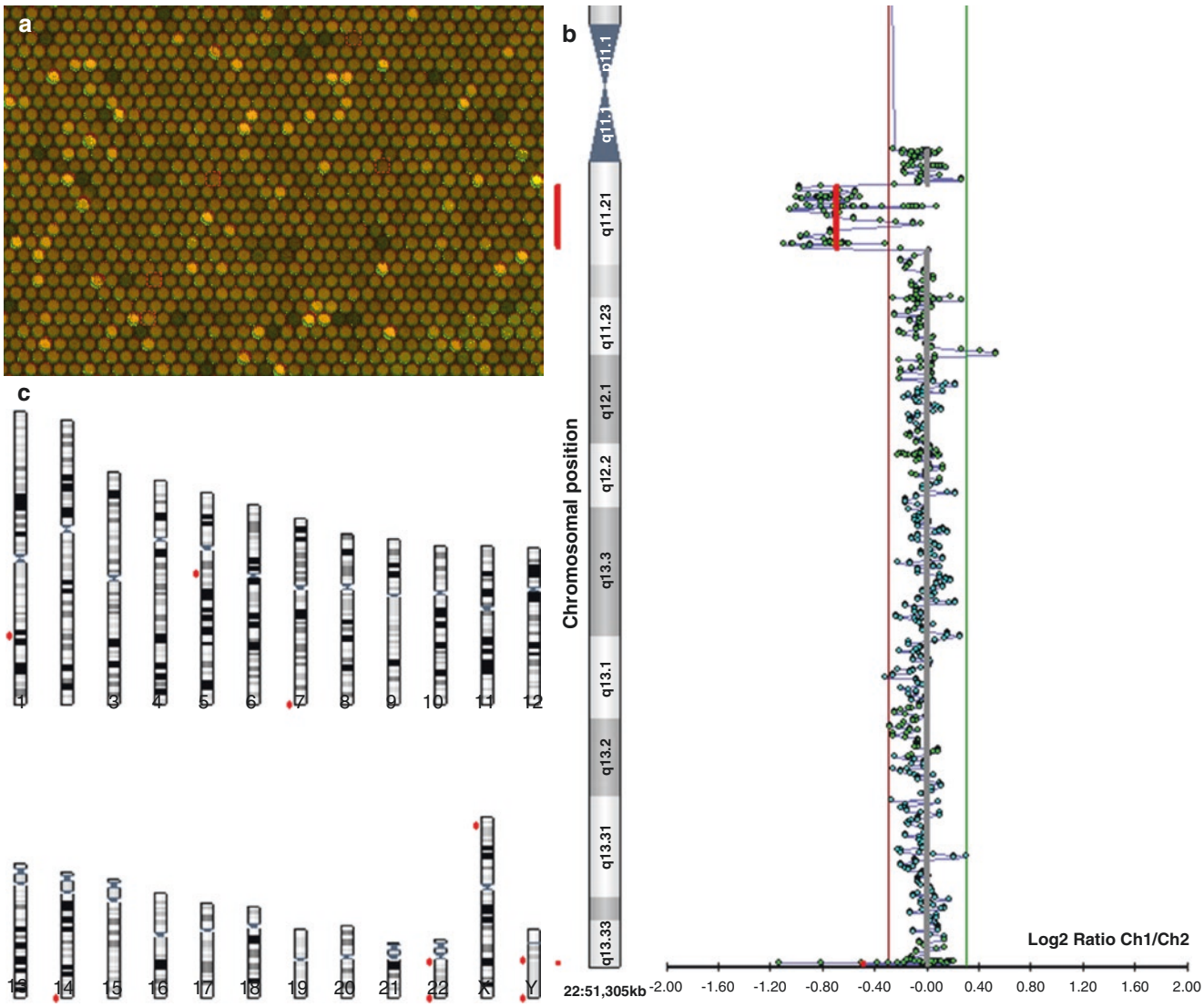
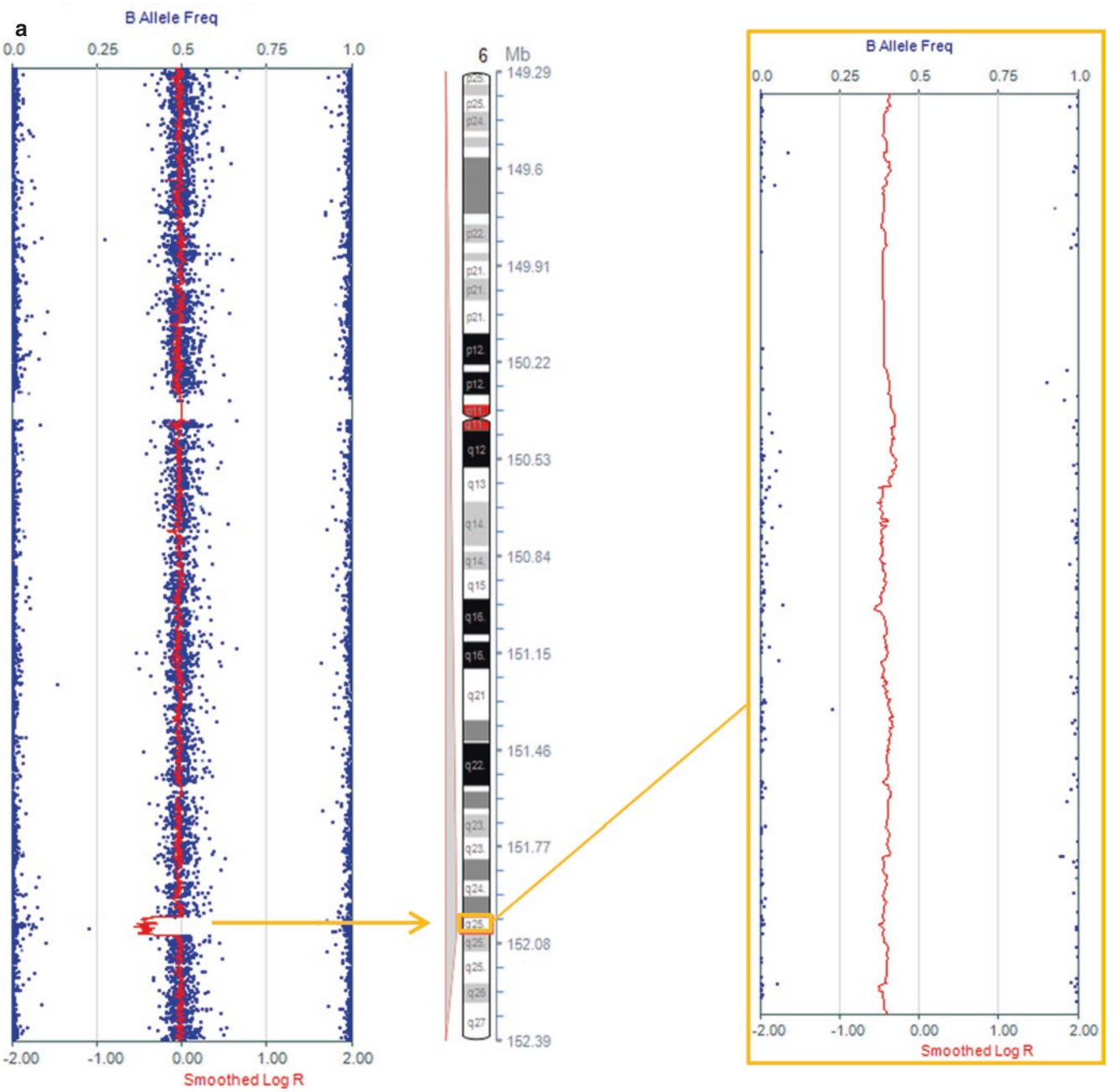


Fig. 3.21 Comparative genomic hybridization (CGH) microarray. (**a**) Fluorescent imaging of CGH microarray chip. *Yellow spots* are the result of equal levels of hybridization between control (*red labeled*) and test DNA (*green labeled*), indicating no copy number variation (CNV). Greater green intensity indicates a relatively greater level of test sample hybridization; i.e., a CNV increase (e.g., from a duplication). Greater *red* intensity indicates a relatively higher level of control DNA hybridization; i.e., a CNV decrease in the test sample compared to control (e.g., from a deletion). Although subtle and not very obvious to the human eye, sophisticated imaging technology is able to discriminate small differences between *red* and *green* intensities, with software indicating (by *red broken line squares*) spots that have a greater red intensity (deletion). (**b**) CGH array readout showing a heterozygous deletion in chr22q11.21, indicated by a cluster with a marked decrease in the log 2 value (< −0.5; highlighted by the *red line*). Classical karyotyping for this child was normal, demonstrating the utility of the higher resolution genetic information obtained by CGH array. (**c**) A virtual karyotype generated from CGH array data in (**b**). Decreased CNVs are indicated by *red dots*, including chr22q11.21 (close to the centromere), associated with DiGeorge syndrome and consistent with the presenting phenotype. Note that not all CNVs are necessarily pathogenic. The region and nature of the CNV must be consistent with the presenting phenotype and currently available evidence of pathogenicity (Figures courtesy of Ms. J. Nicholl, SA Pathology, Australia)

Next-Generation Sequencing (NGS)

The power of **next-generation sequencing** (**NGS**), also known as **massively parallel sequencing** (**MPS**), comes from the ability to quickly and cheaply sequence billions of small fragments of DNA simultaneously (in parallel), combined with powerful, affordable computing for analysis of the large data sets produced (bioinformatics).

Since 2008, the improvement in sequencing technology output and cost has accelerated at a much greater than exponential rate allowing it to be offered clinically to individual patients today. In 2014, the wet lab component of NGS reached a landmark cost of US$1,000 per genome, with data generated in less than a day. This exceptional rate of improvement in mass sequencing can only fully be appreciated when it is compared to the completion of sequencing of the first


a
B Allele Freq
0.0
0.25
0.5
0.75
1.0
6
Mb
149.29
149.6
149.91
150.22
150.53
150.84
151.15
151.46
151.77
152.08
152.39
-2.00
-1.00
0.00
1.00
2.00
Smoothed Log R
B Allele Freq
Smoothed Log R
b
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
zoom in
zoom out
chr6:149,287,703-152,387,664
3,099,962 bp.
enter position, gene symbol or search terms
go
RefSeq Genes

annotated human genome in 2003 (Human Genome Project), costing US$2.7 billion and taking more than a decade.

NGS has the following main steps:

- Fragment whole or part of the genome into smaller fragments.
- Use PCR to amplify and incorporate a common sequence, plus unique identifier sequence (barcode) onto the end of the genomic DNA fragments.
- Bind the amplified fragments via the incorporated common sequence onto microchips.
- Sequence billions of the bound fragments simultaneously by measuring nucleotide incorporation.[3]
- Align all of the sequences from the fragments in software, using a reference human genome for alignment.
- Use filters to determine which changes that are divergent from the reference genome are known or predicted to be important.

An example of the output of an NGS procedure is given in Fig. 3.23. Note, in this case, there are single-nucleotide variations on both alleles, one a substitution and the other a deletion, indicating compound heterozygous mutations. This demonstrates the power of NGS in that a very large number of individual fragments covering this region are sequenced individually, rather than the averaging approach of Sanger sequencing. It is also useful for demonstrating somatic differences present at very low percentage compared to germline tissue.

The number of times a region is individually sequenced is called **coverage depth**, and obviously the larger this number indicating the same sequence variation, the greater the confidence it is a real variation in that individual's DNA.

Limitations of the different systems available for NGS include difficulties in sequencing GC-rich regions and length of reads only in the hundreds of base pairs range, making it difficult to detect **in**sertions or **del**etions (**indels**) greater than approximately 50 bp. **Internal tandem repeats** or **homopolymer repeats** (of the same nucleotide, e.g., CCCCCC) can also cause sequencing problems in NGS.

[3]The two major NGS platforms do this by either imaging fluorescently labeled nucleotide incorporation or measuring pH change associated with hydrogen ion release during DNA polymerization of newly incorporated nucleotides.

There are a range of types of NGS based on how much of the genome is actually sequenced:

- Panel: uses a preamplification step to select for regions of interest (e.g., only exons associated with cardiomyopathy)
- Whole exome sequencing (WES): exons only
- Whole genome sequencing (WGS): the entire genome

It should be noted that the NGS platform is suitable for application to any form of nucleic acid-based sequencing (genome, exome, transcriptome, methylome/epigenome, microbiome).

The post-wet lab component of data analysis to classification of findings and generation of a report is summarized in bioinformatics (page 81).

- **NGS** panel testing is a good test for congenital cardiomyopathies.
- **WES** is a good test for identifying new disease genes or non-classical presentation of a known syndromic condition, where insufficient clinical features have not raised suspicion regarding that syndromic diagnosis.
- **WGS** looms as a good test for almost every diagnostic genetic indication.

Non-invasive Prenatal Testing (NIPT)

Cell-free DNA (**cfDNA**), as it name suggests, is small fragments of DNA (150–200 kb) freely circulating in plasma, no longer associated with its cell of origin. It probably arises from a combination of cell death (i.e., apoptosis) plus extracellular "shedding" from intact cells; however, the mechanisms involved in its production are still far from clear. From 7 weeks gestation, in addition to their own maternal cfDNA, some fetal cells and fetal cfDNA (**cffDNA**) derived from placenta are present in the plasma of pregnant women.

Harnessing the power of massively parallel sequencing to individually sequence billions of fragments of DNA simultaneously, minute amounts of fetal cfDNA can be detected even when it is only a small percentage of the total cfDNA in maternal plasma. By increasing the coverage depth and decreasing the numbers of regions assessed, massively parallel sequencing can theoretically sequence all molecules of cfDNA within a single sample. If there are even small

Fig. 3.22 Single-nucleotide polymorphism (SNP) microarray. (**a**) Heterozygous deletion in chr6q25. Copy number variation (CNV) is indicated by a change in both B allele frequency (<0.5) and smoothed log R (<0) values, shown as a dipping *red line* in both gross (*left*) and fine (*right*) readouts. A virtual karyotype (*center*) indicates the region the deletion is found in chromosome 6 (*orange box*). (**b**) List of known RefSeq genes in the deleted region (*red box*) using UCSC Genome Browser. This includes the gene *TAB2*, associated with heterozygous cardiac development conditions, consistent with the phenotype (Figures courtesy of Ms. F. Norris, Victorian Clinical Genetics Services, Australia)

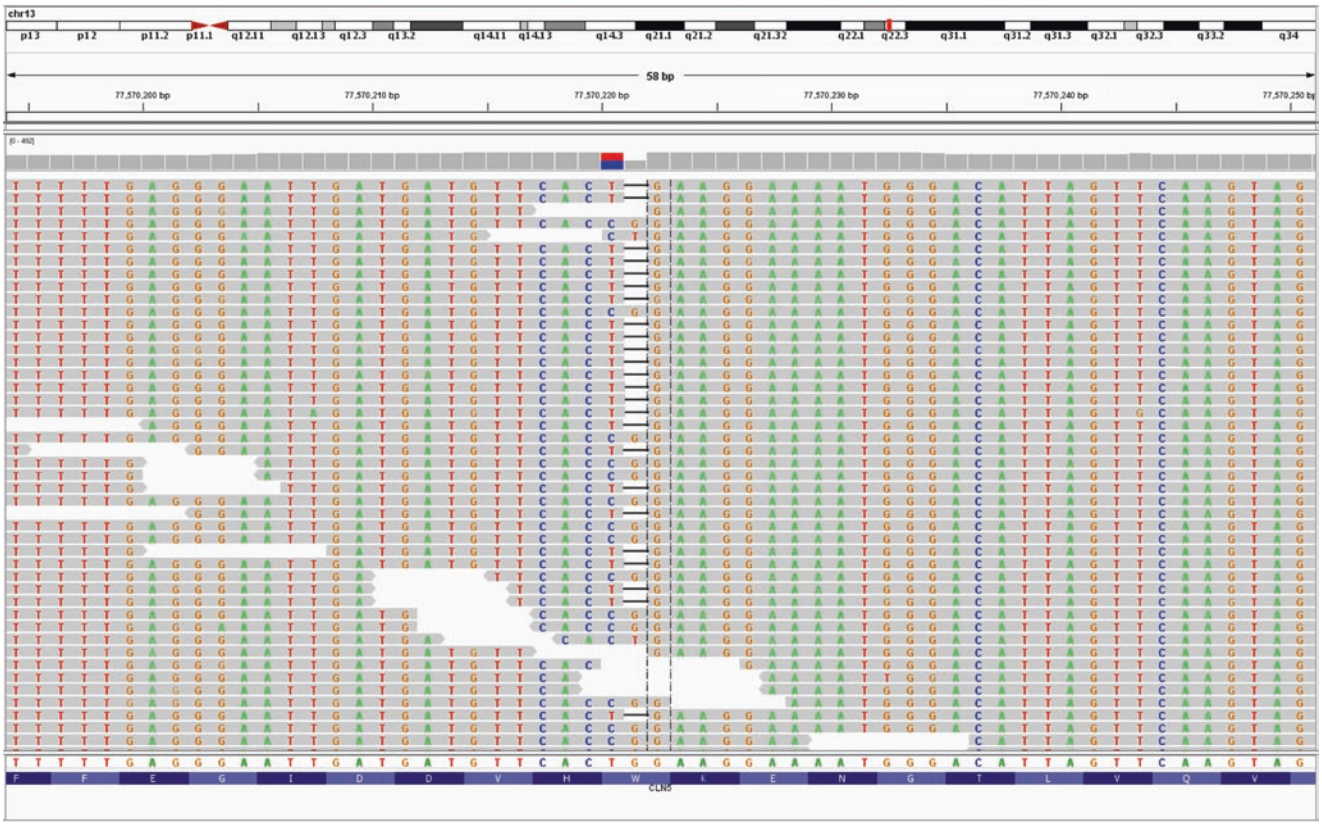


Fig. 3.23 Next-generation sequencing (NGS) of a compound heterozygous mutation. NGS sequences billions of small fragments of DNA in parallel, aligning each individual sequence to a reference human genome via software. In contrast to Sanger sequencing, this produces sequence readout able to show differences down to the level of individual fragments of DNA. Sophisticated bioinformatic pipelines allow the data to be filtered according to many criteria, including quality of sequence, confidence of results, frequency, prevalence, clinical phenotype, and known disease associations. Sequencing occurs in both the forward and reverse directions simultaneously, with even adjacent mutations on separate alleles able to be clearly visualized. Shown is compound heterozygous mutations in the *CLN5* gene, associated with ceroid lipofuscinosis, neuronal, type 5. Gene location is indicated by a *red vertical line* through chr13q22.3 (*top line*) and numerically by genomic coordinates below that. The c.670T>C mutation on one allele is indicated by a color change from *red* to *blue* in the *rectangles* immediately above the sequence data as well as the individual letters of the sequence. Immediately adjacent, the c.671delG deletion mutation on the second allele is denoted by a *white rectangle* above the data, with a *black horizontal line* in the individual sequences (Figure courtesy of Mr K. Brion, SA Pathology, Australia)

amounts of change in the relative quantities of sequence associated with specific chromosomes in the cfDNA, it can indicate aneuploidy. It may also be used for sex determination (detection of any Y chromosome cfDNA will indicate a male fetus, as the maternal cfDNA should have no Y chromosome material). It is also currently used for rhesus D blood grouping. Theoretically, it could be rolled out for detecting any other condition related to copy number variations in specific genes, but currently, there are no commercially available single-gene applications of NIPT.

The main advantage for this technique is its non-invasive nature, compared to other prenatal cytogenetic techniques (CVS and amniocentesis). It can be performed with nothing more invasive than venipuncture for the mother and essentially none of the risk of fetal loss associated with other invasive techniques.

Analysis relies on a statistical number crunching exercise. For example, if 6 % of total cfDNA is fetal and chromosome 21 (being one of the smaller chromosomes) represents 1.5 % of the DNA in a genome, then a trisomy 21 will increase the amount of chromosome 21 fetal cfDNA in maternal plasma by 0.125 % to give 1.625 % of total fetal cfDNA, indicating Down syndrome. Currently it is recommended that positive NIPT tests be confirmed with traditional karyotyping, particularly as the source of fetal cfDNA is placental and therefore a healthy fetus with placental mosaicism would be incorrectly classified using NIPT results alone.

The negative predictive value of the test, with a high-risk antenatal serum screen, is of the order of 99 %. Therefore it is certainly likely to be useful in decreasing the number of cases that progress to invasive sampling for classical karyotyping.

NIPT technology is rapidly gaining popularity, not the least because of its non-invasive attributes. It is also being heavily marketed by a handful of companies offering competing platforms. Initial prospective population study results have been promising [45, 46]; however, it still requires large-scale empirical data from screening populations to validate its reliability. It is advised to be careful of the commercial literature

claims for specificity and sensitivity for NIPT as they mostly utilize retrospective cohorts, with the selection bias inherent from high-risk populations. Data on the influence of factors such as weight, age, ethnicity, and previous pregnancies on cfDNA quality and quantity, especially for false-positive rates, are still to be established in large screening populations [47, 48]. Fetal cfDNA less than 4 % of the total cfDNA in maternal plasma is not sufficient for a reliable result. This technology is not going to completely replace antenatal serum screening or karyotyping, but the proportion of prenatal screening it is utilized for is set to continue to increase.

- **NIPT** is a good screening test for trisomies 13, 18, and 21; monosomy X; and sex determination for X-linked disorders, but a positive result requires confirmation by an invasive test.

Bioinformatics

NGS has reached its current level of relatively wide availability due to both rapid advances in the core sequencing technology and parallel development of analytical tools on very powerful, yet affordable, computing platforms. The latter has pushed the field of bioinformatics to the very prominent position it enjoys today, as the engine behind NGS, deriving clinically significant meaning from the vast data generated by this technology.

The main proprietary NGS technology platforms offer locked-down software analysis tools; however, a very collaborative bioinformatics research community is producing superior and more customisable analysis "pipelines". Just like NGS wet lab technology, the bioinformatics supporting data analysis is evolving rapidly. There is likely to be further improvement in speed and automation of initial raw data filtering protocols as more laboratories adopt this platform for increasing diagnostic uses.

The basics of a bioinformatic analysis pipeline from the filtering stage are illustrated in flow diagram form in Fig. 3.24. The Broad Institute offers a useful set of imaging and analysis tools (the Genome Analysis Toolkit, GATK) that are a good starting point [49]. The basic bioinformatic steps can be summarized as:

- Initial wet lab component for data generation and storage
- Quality assurance checks within data and acquisition machinery
- Alignment of sequence according to the latest **assembly** of the reference human genome (at time of print GRCh38/hg38[4])
- Initial filtering to remove common (non-pathogenic) variants
- Annotation with clinical indication information
- Interrogation of integrity of individual variants
- Comparison to known pathogenic variants
- Pathogenicity prediction algorithms to determine likelihood of a new variant being pathogenic
- Determination on the level of pathogenicity using prescribed guidelines
- Seeking further evidence for pathogenicity using disease databases and biomedical literature
- Reporting as either pathogenic, likely pathogenic, variant of unknown significance, likely benign, or benign according to guidelines [50]

There are various databases to check for common variants (URLs for online tools shown in Fig. 3.24 are listed in references [16, 17, 51–59]).

It should be noted that while the data acquisition component is usually comprehensive, normally only a subset of the genome, associated with the indication, will be probed for analysis.

At this stage, although the current tools are very sophisticated, the bioinformatics field is analogous to early days of personal computing where protocols and code and those literate with them dominate. It is hoped and likely that the field will emulate the progression of personal computing with evolution into more user-friendly interfaces and intuitive manipulation tools for analysis, with some offerings to lay consumers already on presentation in direct-to-consumer genetic tests. For NGS, the term ***in silico*** is currently used to describe computer-based analysis or simulation, particularly with regard to prediction of pathogenic variants.

The quality of the bioinformatic pipeline in generating clinically meaningful results is still and always likely to be highly dependent on the quality of the clinical information provided. Bioinformaticians get even more upset with blank clinical indication fields than pathologists or medical scientists, as despite what they say, it is not all about the data. Continuing development of nomenclature standards, including more defined categorization of levels of evidence, will also enhance this process.

Future of Genetic Testing

The NGS platform is suitable for any nucleic acid-based tests, including analysis of the epigenome (miRNA, lncDNA, CpG methylation), an emerging field likely to continue making inroads into the diagnostic realm. There is also recent evidence that somatic changes may be responsible for some congenital diseases, e.g., in brain development [60]. There is already much experience using NGS for somatic analysis in

[4] www.ensembl.org/Homo_sapiens/Info/Index

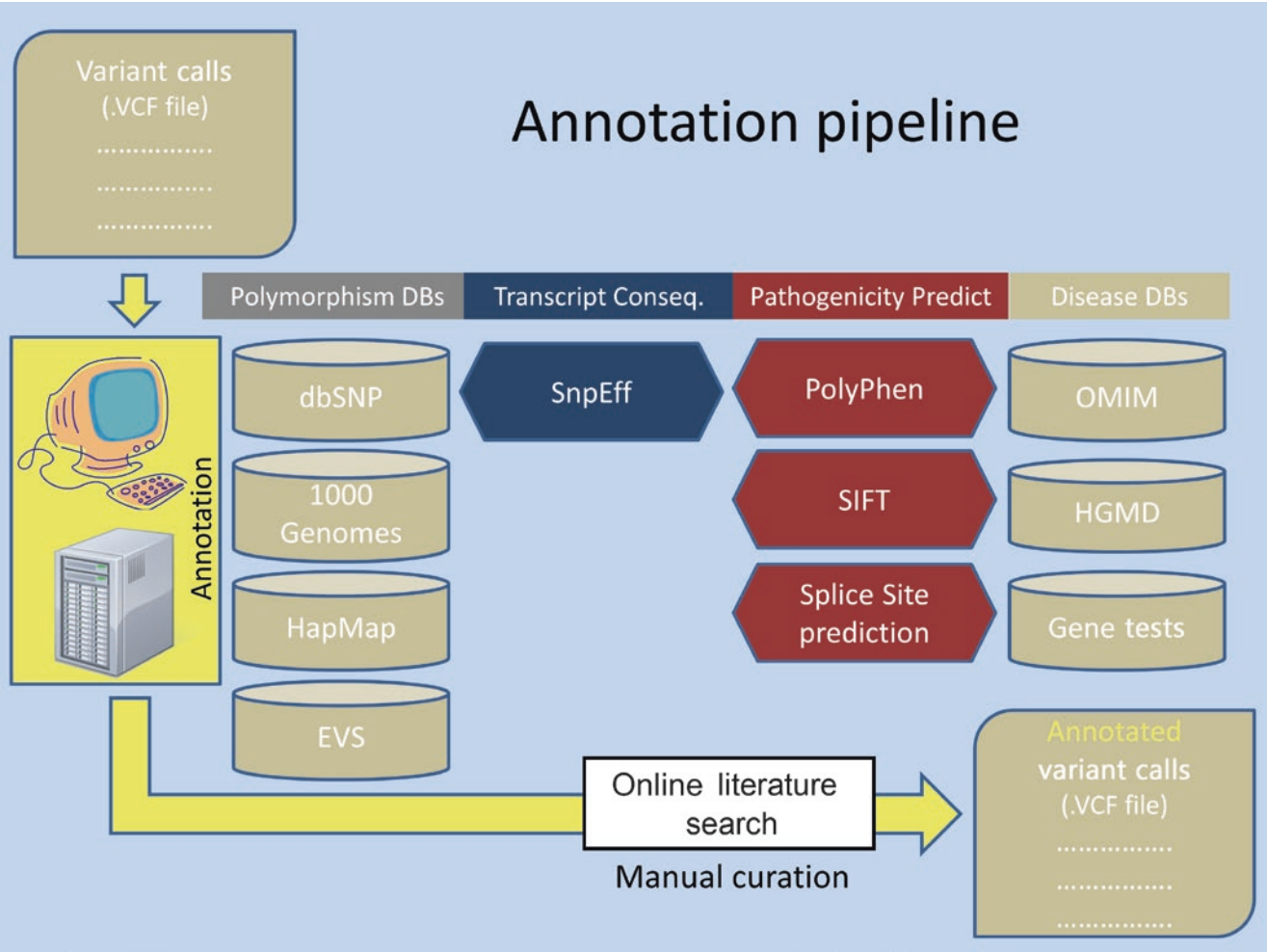


Fig. 3.24 Example of a bioinformatic annotation pipeline for next-generation sequencing (NGS). The annotation process combines polymorphism and disease databases (DBs) with transcription consequence and pathogenicity prediction tools, plus manual curation by traditional literature searches (Figure courtesy of Dr K. Kassahn, SA Pathology, Australia. Web pages for all of the resources in this figure are included in Refs. [16, 17, 51–59])

the cancer field. It is likely that we will see continuing convergence of germline and somatic genetics, alongside the epigenetic revolution, leading to further large paradigm shifts about the genetics of very early development.

NGS is likely to find even further applications not yet identified, as the wet lab price plus automation and experience in the bioinformatics component continue to improve. Given the vast amounts of data generated and computing power required, the whole field is ripe to the advantages and caveats associated with cloud computing [61]. NGS is also increasingly being used in microbiology for rapid characterization of infectious organisms as well as native flora (microbiome), with much interest in marker profiles of both future good and poor health.

Immunogenomics (genetic response of the immune system) and pharmacogenomics (genetic profiling of drug response) are also areas that promise much, generating much interest and research, with both reliant on NGS technologies for data acquisition. There may come a time when NGS-based genetic profiling of a range of factors in one individual from a single sample is used as a diagnostic grab-all in an acute setting; filtered, sifted, and reinterrogated as the differential diagnoses are refined.

There is much hope for these new genetic-based diagnostic technologies, with many blue sky promises and much marketing hype behind them. However, the value of the data they generate will continue to be determined by the quality of clinical description—human factors that are unlikely to be superseded by technology any time soon.

There are many ethical considerations that will emerge from the new genetic testing regimes, and a variety of guidelines and laws are likely to be created across many different professional and societal jurisdictions [62]. Many new questions have emerged from recent genetic testing technologies,

but for many individual patients and families, it has already given answers previously not able to be found by other diagnostic odysseys.

References

1. HGNC. HUGO Gene Nomenclature Committee; Available from http://www.genenames.org/.
2. Anonymous. DNA replication: how genetic information is passed on, 3D animation with narration. The Cold Spring Harbor Laboratory. Available from www.dnalc.org/view/15530-DNA-replication-how-genetic-information-is-passed-on-3D-animation-with-narration.html.
3. Anonymous. Codons and amino acids. Hum Genome Var Soc. Available from www.hgvs.org/mutnomen/codon.html. 4 Jan 2015.
4. Ward AJ, Cooper TA. The pathobiology of splicing. J Pathol. 2010;220:152–63.
5. Trent RJ. Molecular medicine: genomics to personalized healthcare. 4th ed. Amsterdam: Elsevier/AP; 2012.
6. Anonymous. Genbank. National Institutes of Health. Available from ncbi.nlm.nih.gov/genbank. 4 Jan 2015.
7. Anonymous. Ensembl European Molecular Biology Laboratory. Available from http://www.ensembl.org. 4 Jan 2015.
8. Anonymous. DNA Databank of Japan. Available from http://www.ddbj.nig.ac.jp. 4 Jan 2015.
9. Anonymous. International Nucleotide Sequence Database Collaboration. Available from www.insdc.org. 4 Jan 2015.
10. Anonymous. Ucsc Genome Bioinformatics. University of California, Santa Cruz. Available from http://genome.ucsc.edu. 4 Jan 2015.
11. Anonymous. Vega genome browser. Wellcome Trust Sanger Institute. Available from http://vega.sanger.ac.uk/index.html. 4 Jan 2015.
12. Anonymous. The Atlas For Genetics and Cytogenetics in Oncology and Haematology. Available from http://atlasgeneticsoncology.org/Educ/NomMutID30067ES.html. 4 Jan 2015.
13. Wain HMBE, Lovering RC, Lush MJ, Wright MW, Povey S, HUGO Gene Nomenclature Committee. Guidelines for human gene nomenclature. Genomics. 2002;79:464–70.
14. Anonymous. Quick reference – simple examples description of sequence changes. Hum Genome Var Soc. Available from www.hgvs.org/mutnomen/quickref.html. Updated Jan 2011, 4 Jan 2015.
15. Anonymous. Recommendations for the description of sequence variants. Hum Genome Var Soc. Available from www.hgvs.org/mutnomen/recs.html. Updated 8 Jul 2013, 4 Jan 2015.
16. Anonymous. dbSNP short genetic variations. NCBI. Available from www.ncbi.nlm.nih.gov/projects/SNP. 4 Jan 2015.
17. Anonymous. Online Mendelian Inheritance in Man® (OMIM). An online catalog of human genes and genetic disorders OMIM. Available from www.omim.org. 4 Jan 2015.
18. Anonymous. dbGaP. National Center for Biotechnology Information (NCBI). Available from www.ncbi.nlm.nih.gov/gap. 4 Jan 2015.
19. Köhler SDS, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. 2014. Available from: www.human-phenotype-ontology.org. 4 Jan 2015.
20. Groth P KI, Kirov I, Traikov B, Leser U, Weiss B. PhenomicDB. Available from: www.phenomicdb.de. Updated 19 Sept 2014, 4 Jan 2015.
21. Anonymous. Enzyme nomenclature database. International Union of Biochemistry and Molecular Biology (IUBMB) Nomenclature. Available from http://expasy.org/enzyme. 4 Jan 2015.
22. Anonymous. Roadmap Epigenomics Project. NIH Roadmap Epigenomics Mapping Consortium. Available from www.roadmapepigenomics.org/overview. 4 Jan 2015.
23. KEGG: Kyoto Encyclopedia of Genes and Genomes. Kanehisa Laboratories; a useful compilation of biochemical pathways and biological basis of drug interactions. Available from: www.genome.jp/kegg.4 Jan 2015.
24. Wright MW. A short guide to long non-coding RNA gene nomenclature. Hum Genomics. 2014;8:7.
25. Anonymous. Catalogue of somatic mutations in cancer (COSMIC). Wellcome Trust Sanger Institute. Available from http://cancer.sanger.ac.uk/cosmic. 4 Jan 2015.
26. International Standing Committee on Human Cytogenetic Nomenclature, Shaffer LG, McGowan-Jordan J, Schmid M. ISCN 2013: an international system for human cytogenetic nomenclature (2013). Basel: Karger; 2013.
27. Anonymous. Genetics Home Reference. U.S. National Library of Medicine. Available from http://ghr.nlm.nih.gov. Updated 12 Jan 2015, 15 Jan 2015.
28. Anonymous. Scitable. Nat Educ. Available from www.nature.com/scitable. 4 Jan 2015.
29. Anonymous. Fact sheets about genetic and genomic science. National Human Genome Research Institute (NHGRI). Available from www.genome.gov/10000202. Updated 7 Aug 2014, 4 Jan 2015.
30. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. Genome Med [Rev]. 2014;6:60.
31. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. 2013;152:1308–23.
32. Kalish JM, Jiang C, Bartolomei MS. Epigenetics and imprinting in human disease. Int J Dev Biol. 2014;58:291–8.
33. Anonymous. Video animation: RNA interference. Nature Publishing Group; animation of RNAi mechanism. Available from www.nature.com/nrg/multimedia/rnai/animation/index.html. 4 Jan 2015.
34. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. Cold Spring Harb Perspect Biol. 2014;6.pii: a018382.
35. Anonymous. Long Noncoding Rna Database v2.0: the reference database for functional long noncoding RNAs. Garvan Institute. Available from www.lncrnadb.org. Updated 3 Sept 2014, 4 Jan 2015.
36. Anonymous. International Human Epigenome Consortium. Available from www.ihec-epigenomes.org. 4 Jan 2015.
37. Anonymous. HEP: Human Epigenome Project. Available from www.epigenome.org. 4 Jan 2015.
38. Bae JB. Perspectives of international human epigenome consortium. Genomics Inf. 2013;11:7–14.
39. Schmidt EE PO, Buhlmann S, Kerr G, Horn T, Boutros M. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes. Available from www.genomernai.org. 4 Jan 2015.
40. Eijk-Van Os PG, Schouten JP. Multiplex ligation-dependent probe amplification (MLPA®) for the detection of copy number variation in genomic sequences. Methods Mol Biol. 2011;688:97–126.
41. Sørensen KM, Agergaard P, Olesen C, Andersen PS, Larsen LA, Ostergaard JR, et al. Detecting 22q11.2 deletions by use of multiplex ligation-dependent probe amplification on DNA from neonatal dried blood spot samples. J Mol Diagn. 2010;12:147–51.
42. Sørensen KM, El-Segaier M, Fernlund E, Errami A, Bouvagnet P, Nehme N, et al. Screening of congenital heart disease patients using multiplex ligation-dependent probe amplification: early diagnosis of syndromic patients. Am J Med Genet A. 2012;158A:720–5.
43. Furtado LV, Paxton CN, Jama MA, Tripp SR, Wilson AR, Lyon E, et al. Diagnostic utility of microsatellite genotyping for molar pregnancy testing. Arch Pathol Lab Med. 2013;137:55–63.
44. Manning M, Hudgins L, Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. Genet Med. 2010;12:742–5.
45. Bianchi DW, Parker RL, Wentworth J, Madankumar R, Saffer C, Das AF, et al. DNA sequencing versus standard prenatal aneuploidy screening. N Engl J Med. 2014;370:799–808.

46. Hudecova I, Sahota D, Heung MM, Jin Y, Lee WS, Leung TY, et al. Maternal plasma fetal DNA fractions in pregnancies with low and high risks for fetal chromosomal aneuploidies. PLoS One. 2014;9:e88484.
47. Liao C, Yin AH, Peng CF, Fu F, Yang JX, Li R, et al. Noninvasive prenatal diagnosis of common aneuploidies by semiconductor sequencing. Proc Natl Acad Sci U S A. 2014;111:7415–20.
48. Wang JC, Sahoo T, Schonberg S, Kopita KA, Ross L, Patek K, et al. Discordant noninvasive prenatal testing and cytogenetic results: a study of 109 consecutive cases. 2015;17:234–6.
49. Anonymous. Genome Analysis Toolkit (GATK). The Broad Institute. Available from www.broadinstitute.org/gatk. 4 Jan 2015.
50. Wallis YPS, McAnulty C, Bodmer D, Sistermans E, Robertson K, Moore D, et al. Practice guidelines for the evaluation of pathogenicity and the reporting of sequence variants in clinical molecular medicine. Association for Clinical Genetic Science (ACGS), Dutch Society of Clinical Laboratory Specialists (VKGL) http://www.acgs.uk.com/media/774853/evaluation_and_reporting_of_sequence_variants_bpgs_june_2013_-_finalpdf.pdf. 2013.
51. Anonymous. 1000 Genomes A Deep Catalog of Human Genetic Variation. The 1000 Genomes Project. Available from www.1000genomes.org. 4 Jan 2015.
52. Anonymous. 1000 Genomes Browser. National Center for Biotechnology Information (NCBI). Available from www.ncbi.nlm.nih.gov/variation/tools/1000genomes. 4 Jan 2015.
53. Anonymous. International HapMap Project. Available from http://hapmap.ncbi.nlm.nih.gov. 4 Jan 2015.
54. Anonymous. Exome Variant Server. NHLBI Exome Sequencing Project (ESP). Available from http://evs.gs.washington.edu/EVS. 4 Jan 2015.
55. Anonymous. SnpEff. Available from http://snpeff.sourceforge.net/SnpEff_manual.html. 4 Jan 2015.
56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. PolyPhen-2 prediction of functional effects of human nsSNPs. Harvard; Available from http://genetics.bwh.harvard.edu/pph2. 4 Jan 2015.
57. Anonymous. SIFT. J. Craig Venter Institute. Available from http://sift.jcvi.org. 4 Jan 2015.
58. Anonymous. The Human Gene Mutation Database. The Institute of Medical Genetics in Cardiff. Available from www.hgmd.cf.ac.uk. 4 Jan 2015.
59. Anonymous. GeneTests™. Available from www.genetests.org. Updated 15 Jan 2015, 4 Jan 2015.
60. Jamuar SS, Lam AT, Kircher M, D'Gama AM, Wang J, Barry BJ, et al. Somatic mutations in cerebral cortical malformations. N Engl J Med. 2014;371:733–43.
61. Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundquist A, English A, et al. Launching genomics into the cloud: deployment of mercury, a next generation sequence analysis pipeline. BMC Bioinf. 2014;15:30.
62. Lohmann K, Klein C. Next generation sequencing and the future of genetic diagnosis. Neurother: J Am Soc Exp NeuroTher. 2014;11:699–707.