Chromosome identification methods

Tsvi Benson-Tilsen

29 December 2025

Contents

Abstract **2**

Annotated table of contents **3**

Disclaimers **3**

Context **3**

Synopsis and takeaways **5**

Generalities about chromosome identification **9**
Definition of chromosome identification . . . 9
Number vs. homolog identification . . . 10
Crossover identification . . . 12
Setwise identification . . . 12
More clarifications . . . 14
Context for chromosome identification . . . 14
Why chromosome identification is hard . . . 14
What are chromosomes? . . . 15
Cell culturing vs. isolating-ensembling methods . . . 16
Direct and indirect chromosome identification . . . 17
Indirect chromosome identification . . . 18
Destructive identification and complementary identification . . . 19
Indirect identification requires origin indexing . . . 21
Combining multiple identification methods . . . 22
Combining homolog and number identification . . . 22
Screening and confirmation tests . . . 23
Jointly discriminative feature sets . . . 23
Ground truth tests . . . 25
Isolating-ensembling methods require high-confidence number identification . . . 26
Illustration of needed confidence . . . 26
Setwise calls vs. uncertain calls . . . 27
The full question of the power of an identification method . . . 29
Confidence of intact DNA . . . 29

Tables of methods **30**
Criteria for identification methods . . . 30
Table of most methods . . . 31
Table of best methods by purpose . . . 34

Direct chromosome identification methods **35**
Summary of direct identification methods . . . 35
Light microscopy . . . 39
Banded chromosomes . . . 39
Length and centromere location . . . 40
Speculations about visual identification . . . 44
Staining . . . 45

How staining works . . . 45
Staining for chromosome identification . . . 47
Immunostaining chromosome structural proteins . . . 48
MicroSort . . . 51
Hybridization methods . . . 52
Fluorescence in situ hybridization . . . 54
CRISPR-dCas9 systems . . . 58
Other hybridization methods . . . 59
Standard genotyping . . . 60
Bulk methods . . . 61
Bulk methods in general . . . 61
Centrifugation . . . 61
Flow cytometry . . . 64
Electrokinesis . . . 66
Types of electrokinesis . . . 67
Electrophoresis . . . 69
Dielectrophoresis . . . 72
Hydrodynamic methods . . . 73
Raman spectroscopy . . . 77
Acoustic methods . . . 80
Acoustic force spectroscopy . . . 81
Scanning acoustic microscopy . . . 82
Acoustic sorting . . . 86
Some more identification methods . . . 88
Electron microscopy . . . 88
Atomic force microscopy . . . 89
Detecting allelic proteins or RNAs . . . 91
Even more direct methods . . . 91

Indirect chromosome identification methods **92**
Sequencing a cell culture . . . 93
Diploid culture . . . 93
Haploid culture . . . 93
Sequencing post-meiotic RNA . . . 95
The idea of RNA-based genotyping . . . 95
The problem of spermatid RNA sharing . . . 97
RNA coverage of the genome . . . 98
Proof of concept . . . 99
RNA-based genotyping for oocytes? . . . 101
Complementary identification . . . 101
Chromosome-wise complementary identification . . . 102
Meiotic complementary identification . . . 103
Targeted DNA operations . . . 108
Targeted complementary elimination . . . 109

Acknowledgements **109**

Abstract

Chromosome selection is a hypothetical technology that assembles the genome of a new living cell out of whole chromosomes taken from multiple source cells. To do chromosome selection, you need a method for chromosome identification—distinguishing between chromosomes by number, and ideally also by allele content. This article investigates methods for chromosome identification. It seems that existing methods are subject to a tradeoff where they either destroy or damage the chromosomes they measure, or else they fail to confidently identify chromosomes. A paradigm for non-destructive high-confidence chromosome identification is proposed, based on the idea of complementary identification. The idea is to isolate a single chromosome taken from a single cell, destructively identify all the remaining chromosomes from that cell, and thus infer the identity of the preserved chromosome. The overall aim is to eventually develop a non-destructive, low-cost, accurate way to identify single chromosomes, to apply as part of a chromosome

selection protocol.

Annotated table of contents

If you're looking for a short thing to read, look at the subsections "Context" and "Synopsis and takeaways".

- Abstract. Very short summary.
- Annotated table of contents. (You are here.)
- Disclaimers. Some context about me and this article.
- Context. The significance of chromosome identification in the context of chromosome selection and reprogenetics.
- Synopsis and takeaways. A bird's-eye view of what this article says. Explains some tradeoffs involved in chromosome identification, and briefly introduces the idea of indirect identification to address those tradeoffs.
- Generalities about chromosome identification. Several theoretical points about what chromosome identification is, its various forms, and constraints that apply to it.
- Tables of methods. Gives a list of criteria by which to judge chromosome identification methods, and then judges identification methods by those criteria, and also gives guesses about the best methods for various purposes.
- Direct chromosome identification methods. A big list of methods that try to identify a chromosome by directly interacting with it.
- Indirect chromosome identification methods. A list of methods that try to identify a chromosome indirectly, by directly interacting with some material that is informationally entangled with the chromosome in question.
- Acknowledgements. Thanks.

Disclaimers

(Skippable.)

This is a theoretical paper. I haven't checked any conclusions directly with experiments; rather, this is a review paper that analyzes a bunch of other people's experiments and makes speculative guesses. So it should be taken as preliminary.

I'm not a trained biologist and I haven't worked in a wetlab, so I may have basic misunderstandings. Also, this investigation touches on many different subareas, and includes some speculative ideas. Finally, a lot of important information is available but only at a significant cost of time and effort (e.g. getting good all-inclusive price and time estimates for various protocols), so I neglected many such questions, and instead just make guesses; I tried to indicate when I was doing so, at least with a question mark, but I may have failed to always do so. For these reasons, there are likely to be several mistakes and omissions. I'd be happy to hear about those, especially ones that affect any of the major conclusions. This is a sort of checkpoint: research notes as part of an ongoing investigation, in which I intend to continue checking against expert knowledge.

The technical motivations for some choices made in this article will hopefully be made more clear by future articles.

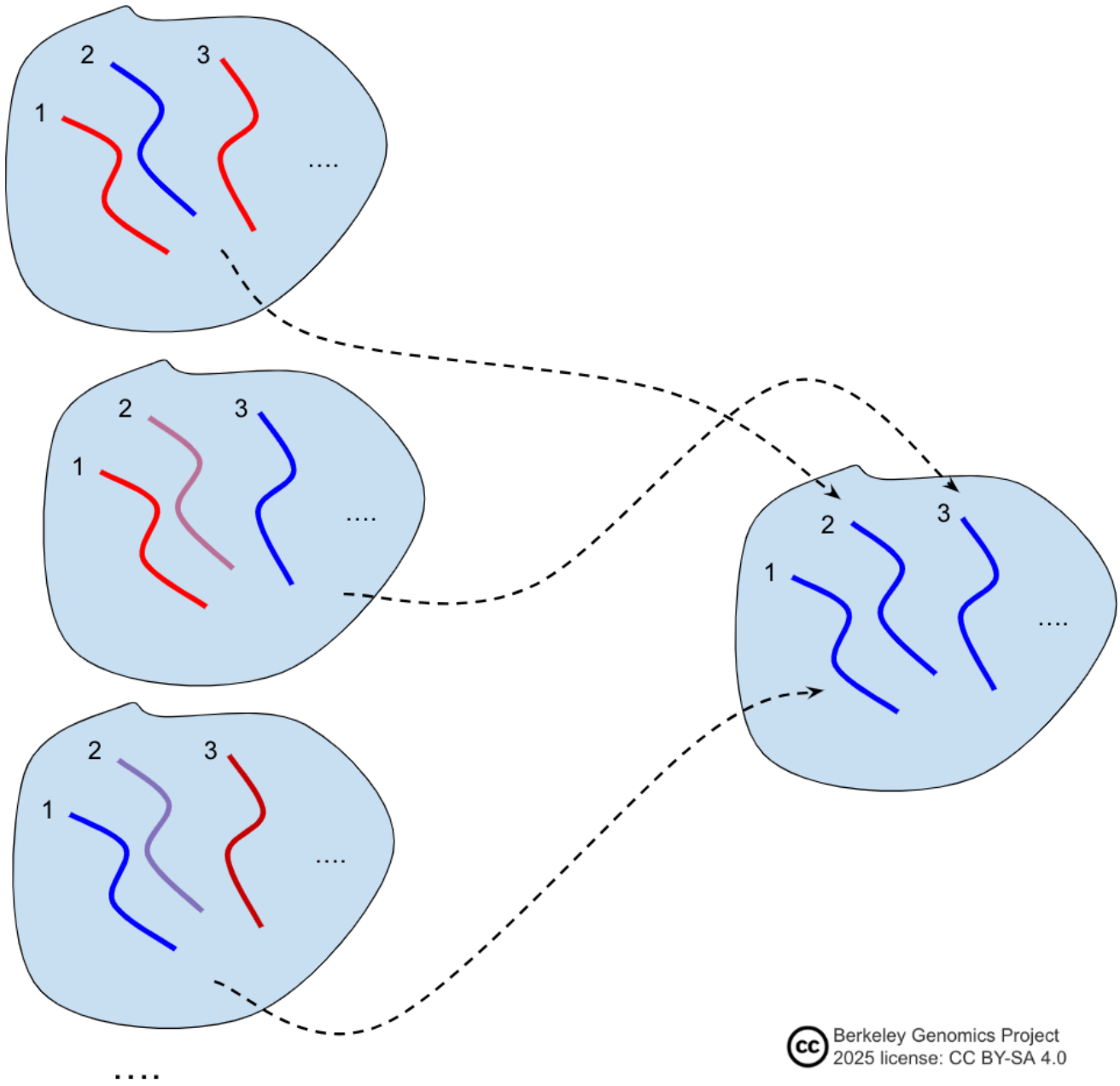
If I were doing this again, I'd spend a bit more effort discussing the current understanding of the physical structure and mechanical properties of metaphase chromosomes and spermatozoon chromatin, which has been somewhat revised and clarified in the past 15 or so years.

[LLM] This manuscript includes figures reproduced from previously published works for the purposes of scholarly analysis, comparison, and critique. All such figures remain the property of their respective copyright holders and are reproduced here with full attribution. No claim of ownership or relicensing is made.

Context

Reprogenetics is biotechnology to empower parents to make genomic choices on behalf of their future children. One key operation that's needed for reprogenetics is genomic vectoring: creating a cell with a genome that's been modified in some specific direction.

Chromosome selection is one possible genomic vectoring method. It could be fairly powerful if applied to sperm chromosomes or applied to multiple donors. The basic idea is to take several starting cells, select one or more chromosomes from each of those cells, and then put all those chromosomes together into one new cell:



There are three fundamental elements needed to perform chromosome selection:

- *Transmission* and *Exclusion*. Get some chromosomes into the final cell, while excluding some other chromosomes.
- *Targeting*. Differentially apply transmission and exclusion to different chromosomes.

This article deals with the *targeting* element. Future articles will deal with the other elements. Specifically, this article tries to answer the question:

> How can we *identify* chromosomes?

That is, how can we come to know the *number* of one or more chromosomes that we are handling (i.e. is it chromosome 1, or chromosome 2, etc.)? Further, how can we come to know what alleles are contained in the specific chromosome we are handling, among whatever alleles are present among the chromosomes we're selecting from?

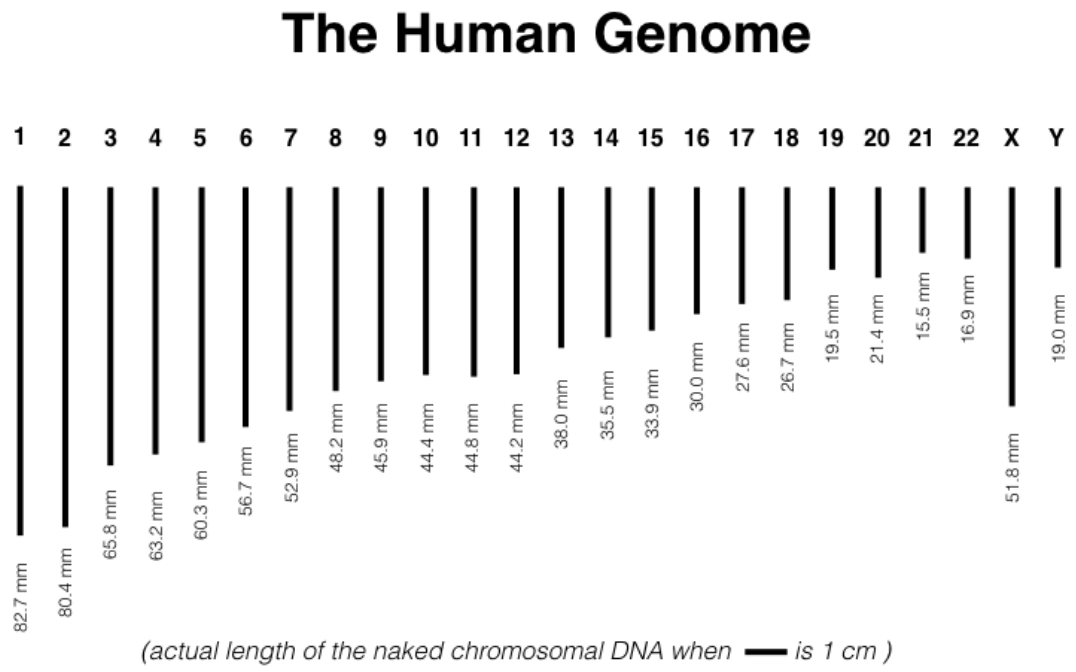
This problem has been approached from many angles. There are several central staples of molecular biology, such as DNA sequencing, karyotyping, flow cytometry, CRISPR-Cas9, and FISH; and there are several speculative attempts to study chromosomes in unusual ways, such as acoustics, laser scattering, hydrodynamic

sorting, and electrokinesis.

This article presents an attempt to sort through these methods and find ones that will work well as part of a chromosome selection method. This goal induces various constraints on methods for chromosome identification; hopefully future articles will further clarify those constraints.

Synopsis and takeaways

A human cell has 46 chromosomes, 2 of each number, with each number (and X and Y) being of different sizes:



(Figure 1.3 from Gallegos (2022)[1]. © publisher)

We want to identify chromosomes. Technically, that means we want to be able to somehow operate differently on chromosomes of different numbers. In practice, for the most part, what we want is to *isolate* one or more chromosomes, and then *learn* what number(s) they are. (If possible, we also want to learn what alleles they carry.)

How do we identify chromosomes? We have to measure them somehow.

There's a tradeoff between different ways of measuring chromosomes: How much *access* do you have to the DNA inside the chromosome? (Chromosomes are not just DNA; they also incorporate many proteins.)

On one extreme, there is, for example, standard DNA sequencing. In this method, you have lots of direct access to the DNA, so you can easily measure it with very high confidence, and learn the number of a chromosome and almost all of the alleles it carries. However, this method is also completely destructive. You strip away all the proteins from the DNA, you disrupt the epigenetic state of the DNA, and you chop up the DNA into tiny little fragments. High DNA access comes with high information, but also comes with high destructiveness.

On the other extreme, there is, for example, standard light microscopy. In this method, you have very little direct access to the chromosome's DNA. You just shine light on the chromosome and see what you can see. This method is not at all destructive; the chromosome's DNA, structural proteins, and epigenetic state are all left perfectly intact. However, this method definitely cannot tell you what alleles the chromosome carries, and may not even be able to distinguish many chromosomes by number. Low DNA access comes with low destructiveness, but also comes with low information.

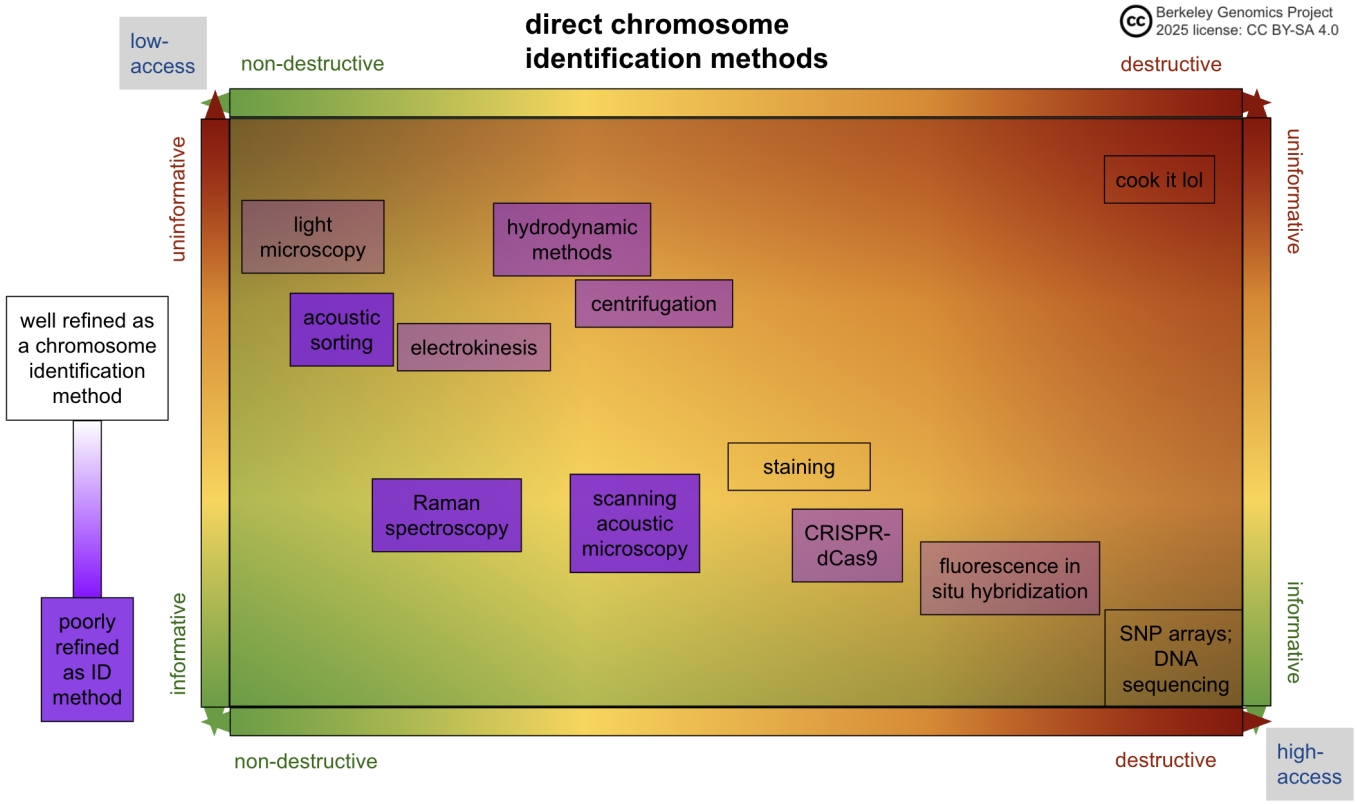
If we're assembling a new cell (for example, to use in place of a sperm), we cannot use chromosomes that we have destroyed. We also (roughly speaking) cannot use a chromosome unless we're confident we know

[1]Gallegos, Maria. Fantastic Genes and Where to Find Them. Updated 2022-09-13. Accessed 16 February 2025. https://bookdown.org/maria_gallegos/where-are-genes-2021/#preface.

what number it is, because we have to be confident that the final cell will be euploid. Are there methods that are non-destructive and also make confident calls about chromosome number?

I don't know of a theoretical reason such a method should not exist. Why not measure physical properties of a chromosome from a distance and infer its number? For example, a single paper from 2006 claimed to use Raman spectroscopy to distinguish with fairly high confidence between human chromosomes 1, 2, and 3, just by bouncing (scattering) a laser off of them[2]. However, all such methods I've looked at are similar, in that they are very poorly refined: they have not been extensively replicated, so they may not work at all, and definitely have not been developed to be easy and reliable.

Therefore, as far as I know, there is currently probably no good way to identify chromosomes by directly measuring them. Every single such method will destroy the chromosome, or will not make confident calls about the chromosome's number, or else has not been well-demonstrated to work. Here's a visual summary of the situation:



(Hi r/ChartCrimes!)

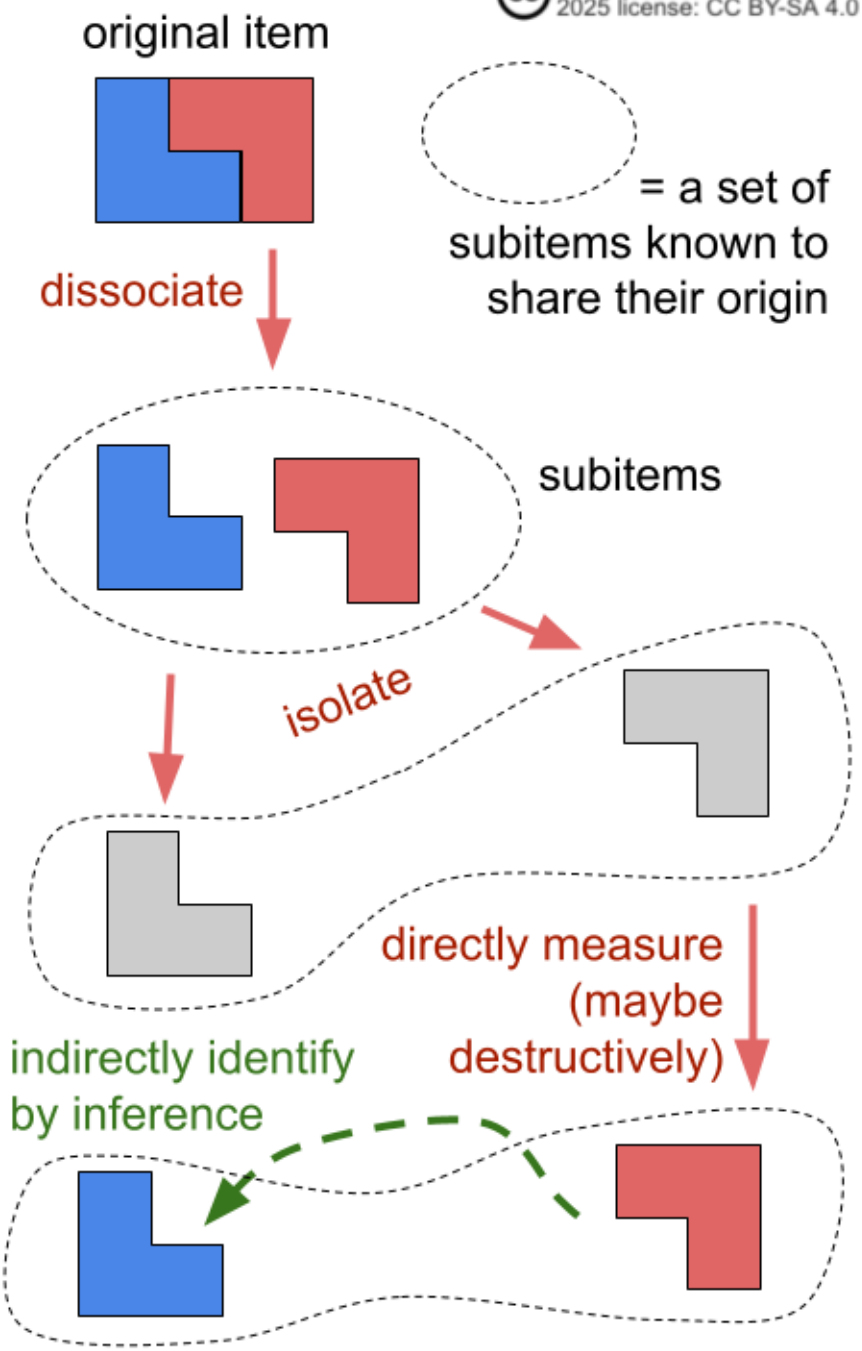
Sidenote: Many readers might wonder: Why not just use standard cell culture sequencing? The reason will be explained more fully in a future article. But basically, the reason is that ensembling a target genome using cell culturing methods (such as MMCT) is likely to be very inconvenient. To avoid that, we want a more reductive mechanical method, an "isolating-ensembling" method, where we isolate single chromosomes, identify them, and then put target chromosomes into a new cell. Isolating-ensembling methods require a way to identify single chromosomes (or small sets of chromosomes); it's not enough to just learn the content of some full euploid genomes, which is all that is offered by cell culture sequencing.

So, if we cannot identify chromosomes by directly measuring them, what to do?

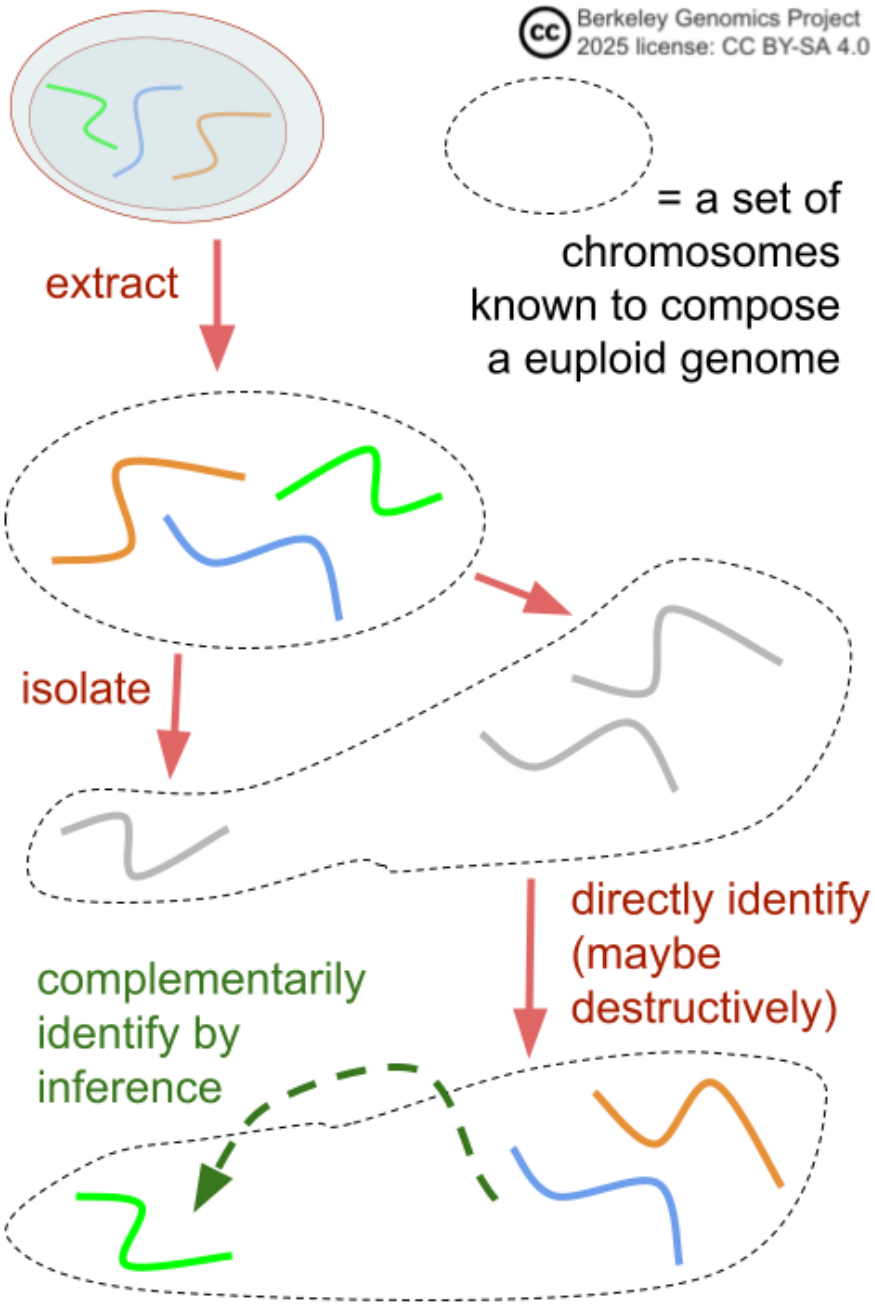
My proposal is to identify chromosomes by *indirectly* measuring them. To indirectly measure a chromosome, we get some material that comes from the same place as the chromosome. We then directly measure that material, and use that measurement to infer something about the chromosome:

[2]Ojeda, Jenifer F., Changan Xie, Yong-Qing Li, Fred E. Bertrand, John Wiley, and Thomas J. McConnell. 'Chromosomal Analysis and Identification Based on Optical Tweezers and Raman Spectroscopy'. Optics Express 14, no. 12 (2006): 5385–93. https://doi.org/10.1364/OE.14.005385

Berkeley Genomics Project
2025 license: CC BY-SA 4.0



A key indirect identification method is complementary chromosome identification. That's where you take a single cell with a known genome, isolate one chromosome, and then sequence the rest of the chromosomes. This tells you the identity of the isolated chromosome, without ever directly measuring that chromosome:



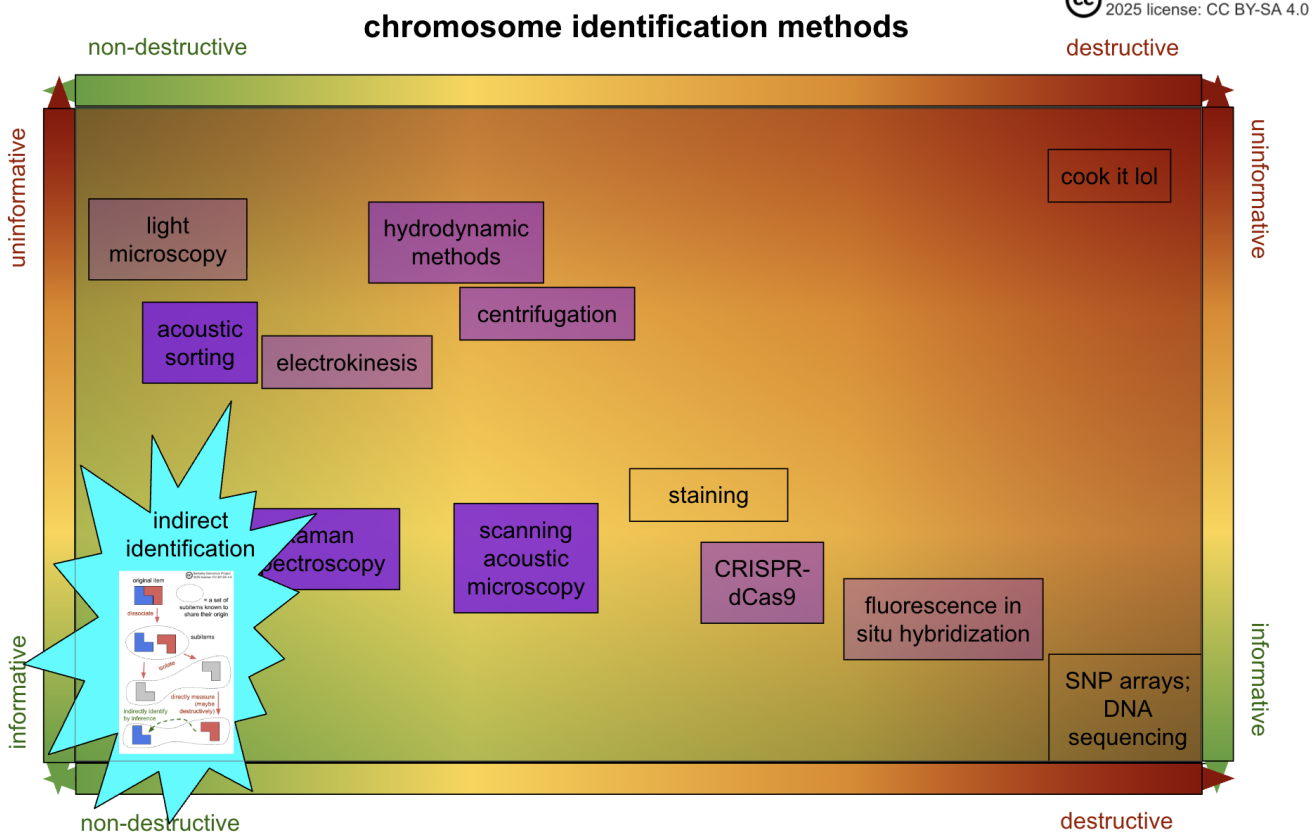
(See the subsection "Chromosome-wise complementary identification".)

Another indirect identification method is single-cell RNA sequencing for sperm. This works by separating out RNAs from a single sperm and sequencing them. It turns out that those RNAs actually tell you which *alleles* are present in that sperm's genome. (See the subsection "Sequencing post-meiotic RNA".) This tells you the *set* of chromosomes you have, including what crossovers happened. (Another way to do this might be to briefly culture the sperm as haploid cells using donor oocytes[3]; see the subsection "Haploid culture".)

By combining complementary chromosome number identification with one of these indirect allele-measuring methods ("setwise homolog identification"), we could in theory isolate a single fully intact chromosome with a confidently, almost completely known genome.

[3]Metacelsus. 'Androgenetic Haploid Selection'. Substack newsletter. De Novo, 16 November 2025. https://denovo.substack.com/p/androgenetic-haploid-selection.

Berkeley Genomics Project
2025 license: CC BY-SA 4.0



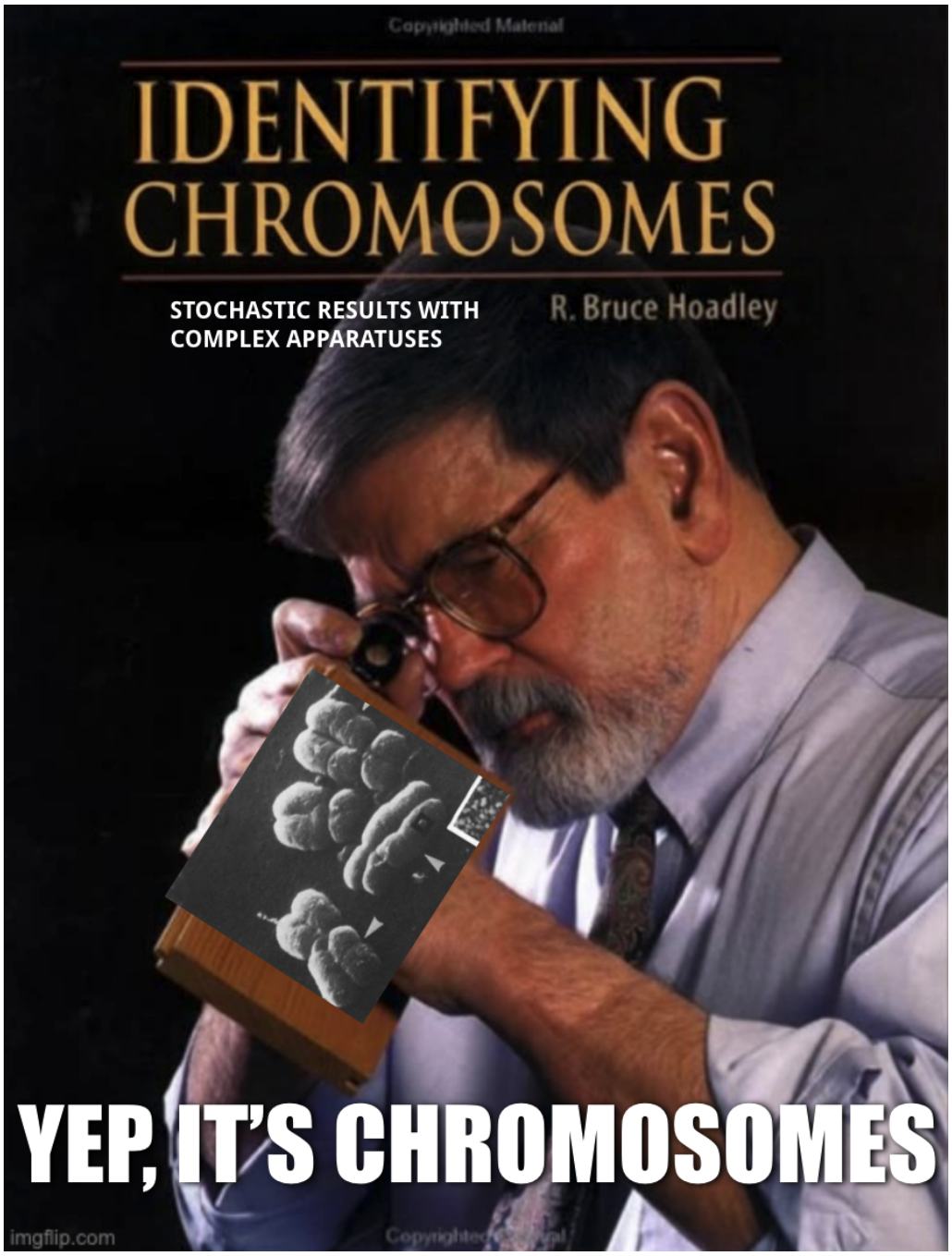
This would be a good solution to chromosome identification. Unfortunately, these methods would be very challenging to actually develop. But, that effort might be worth it, since it seems there are not better chromosome identification methods available. See future articles for discussion of how to implement these methods.

The rest of this article will go into much more detail on many of the above points.

Generalities about chromosome identification

Definition of chromosome identification

> *Chromosome identification* means learning something about the identity of a chromosome—that is, learning something about which chromosome it is out of several possible chromosomes.

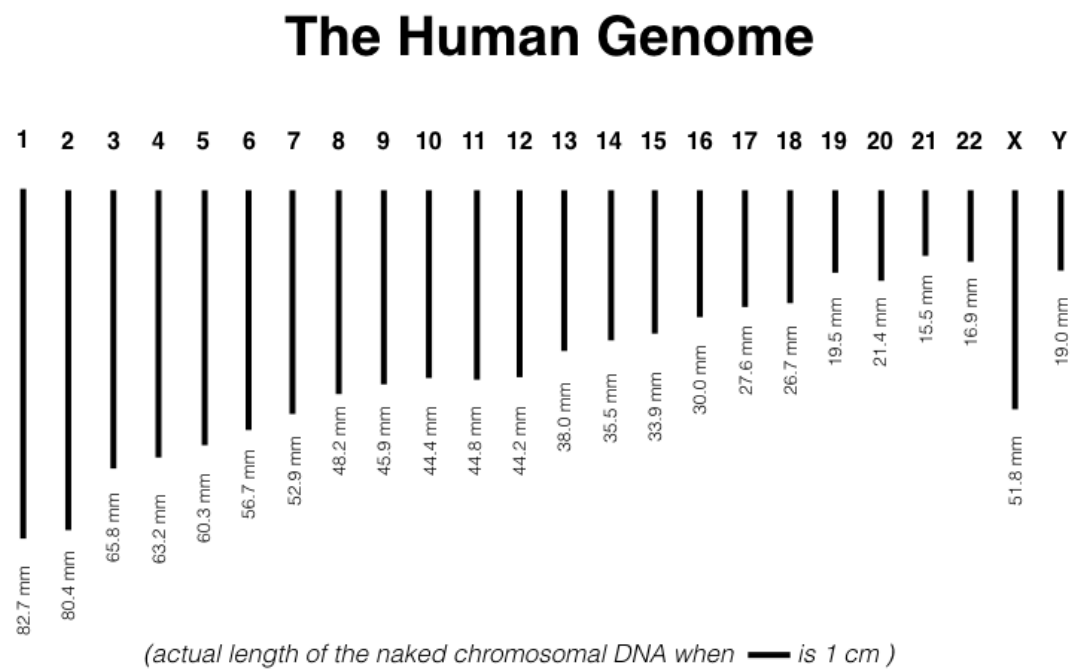


Number vs. homolog identification

> *Chromosome numbering* or *chromosome number identification* means learning the number of a chromosome.

Chromosome numbering is like karyotyping, but for a single chromosome or a small set of chromosomes.

E.g. if it is a human chromosome, numbering means learning whether it is autosome 1, or autosome 2, or autosome 3, or …, or autosome 22, or sex chromosome X, or sex chromosome Y.



(Figure 1.3 from Gallegos (2022)[4]. © publisher)



(Image from https://en.wikipedia.org/wiki/G_banding#/media/File:NHGRI_human_male_karyotype.png. © publisher)

On the other hand:

> *Homolog identification* means learning which of two or more possible homologous chromosomes you have.

In other words, homolog identification distinguishes between two chromosomes that have the same number, but that have different sets of alleles. (Related terminology: "genoinformative".)

E.g., suppose you've extracted the chromosomes from a diploid cell, i.e. a cell that has two copies of chromosome number 1. One copy comes from the cell's organism's father, and the other copy comes from

[4]Gallegos, Maria. Fantastic Genes and Where to Find Them. Updated 2022-09-13. Accessed 16 February 2025. https://bookdown.org/maria_gallegos/where-are-genes-2021/#preface.

the mother. Suppose you've isolated one copy of chromosome 1. But, you still don't know whether it is the paternal or maternal chromosome 1. If you learn whether it is paternal or maternal, you've done homolog identification.

Crossover identification

Another task is to identify which crossover has occurred in a chromosome that came from a gamete. (See "Sensing crossovers".)

In what is perhaps a slight abuse of terminology, we'll also include this task under the title "homolog identification". Some clarifications about this point:

- Throughout this article, we assume, as background context, that we know the full genomes of the sources of the chromosome that we're dealing with.
  - In that context, homolog identification is a reasonable description of the task of identifying crossover locations. There's a fairly small space of possible chromosomes resulting from one or a couple crossovers, so there's a fairly small amount of information needed to identify most of the complete DNA sequence of a given chromosome with fairly high confidence (see "Multiple crossovers"). Therefore, it roughly makes sense to think of the genotyping task as being "chromosome-wise" and fairly discrete—you either know which homolog you have, or you don't know it. In other words, we don't necessarily have to worry about partial knowledge states such as "the first third of the chromosome is paternal, the remaining two thirds are unknown", at least when developing the basics of a chromosome selection protocol. (In practice, if we're doing homolog identification for reprogenetics, we'll have to model uncertainty and partial information.)
  - The practical difference might come down to whether you can label one locus vs. a few loci on a chromosome. A few loci vs. many loci does make that much of a difference, because you just learn very specifically where the crossovers are, rather than roughly; but that doesn't change your belief about most of the genome, which is already pinned down by know approximate crossover locations.
  - If we're extracting chromosomes either from adult somatic cells or from gametes produced by an adult, we assume that we've sequenced the adult's genome. This is a fairly generally applicable assumption, in that it's usually not hard to satisfy the assumption.
  - I haven't much considered cases where it doesn't apply. One possible such case would be chromosome selection applied in combination with iterated editing or iterated recombinant selection, where cell lines are being cultured and chromosomes are being transplanted between cell lines. Due to mutation (intended or unintended), the DNA sequence of the chromosomes in the cell cultures may change significantly. Further, it might be cost-prohibitive to clone and resequence cell lines frequently enough to precisely track all these changes. In this case, it might be more important to track sequence uncertainty quantitatively. For example, one might want to track quantitatively how long (how many cell divisions or other DNA modification events) have occurred since the last sequencing; that corresponds to more and more uncertainty about the exact DNA sequence.
- We ignore minor differences between chromosomes, e.g. due to gene conversion. For reprogenetic purposes, de novo mutations have to be tracked, but segment-swapping between chromosomes only matters when it is large segments that are being swapped.

Setwise identification

An important clarification: chromosome identification can be achieved for sets of chromosomes. That is, you could learn what set of chromosomes you have, *confidently*, without having identified any one specific chromosome.

This can happen, for example, because of correlation between chromosomes. If two chromosomes came from the same euploid haploid cell, you *know* they are not the same number. So, for example, you could identify chromosomes $\{10, 11, 12\}$, so that you're very confident about that being the set you have, without knowing which of the three chromosomes is which of $\{10, 11, 12\}$. (See also the subsection "Setwise calls vs. uncertain calls".)

Setwise identification may or may not be satisfactory on its own, depending on context:

- For example, suppose some method is only able to identify two sets of chromosome numbers, A and B.

If A and B are overlapping, or if they do not cover the whole genome, then this method can't be used to ensemble a full euploid genome, so it is not satisfactory. (It may seem strange to be able to identify B but not the complement of B; but consider for example the possibility of targetedly destroying all chromosomes except for those in B. Now you have B and you know it, but you couldn't necessarily knowingly obtain the complement of B.)

- On the other hand, if a method can identify a family of sets of chromosome numbers that partitions a euploid genome, then it might be satisfactory. As long as the partition-wise calls are correct with high confidence, we can ensemble a euploid genome. The resulting partial chromosome selection method might be underpowered compared to full chromosome selection, but would still be a useful innovation.
- See "Method: Chromosome selection" for analysis of the selection power of CS in general. The power of setwise chromosome selection depends on the size of the sets. If most of the sets are singletons, the selection power is barely affected. For comparison, if almost all of the sets are pairs of chromosomes, then the power would be roughly 0.7 times the power of full chromosome selection. (The variance of the pair is twice the variance of the individuals, roughly; so the standard deviation of the pair is about $\sqrt{2}$ the standard deviation of a singleton; and you have half as many pairs as singletons, giving $\sqrt{2}/2 = 1/\sqrt{2} \approx 0.7$ times as much selection power.)
- Gamete selection is an edge case of chromosome selection which uses haploid-setwise homolog identification.

Examples of setwise identification:

- Ambiguous direct identification.
  - Several direct chromosome identification methods, such as light microscopy or electrokinetic methods, might be able to distinguish chromosomes roughly by large-scale features like length, volume, mass, or centromere position. Most chromosomes are different enough in those features to be distinguishable, but e.g. $\{10, 11, 12\}$ are very similar.
  - This implies that such a method might produce *high-confidence* calls, but with some *unresolved ambiguity*. As discussed above, this could be satisfactory.
- Complementary numbering for droplets.
  - (See the subsection "Chromosome-wise complementary identification".)
  - Suppose you use microfluidics to produce a water droplet containing two chromosomes[5]. You could destructively number all the remaining chromosomes coming from the source cell. Then you would know the set of two chromosome numbers in your droplet, without knowing which is which, and maybe without knowing which homolog you have (if you didn't already know).
  - This might be sufficient on its own. For example, you might then separately produce droplets that together compose the remainder of a full euploid genome. You can merge all those droplets, plus the droplet with the known pair of chromosomes. You've thus produced a genome known to be euploid without ever specifically identifying the number of either of the two chromosomes in that droplet.
  - Setwise complementary numbering could be useful.
    - Example: Suppose that it's easy to do an inexpensive screening test to find droplets with a specific *small set* of chromosomes, but difficult to screen for specific *single* chromosomes. (See the subsection "Screening and confirmation tests".)
    - E.g. if the screen is visually estimating chromosome size, then the screen leaves ambiguity between similar-sized chromosomes.
    - Here, it might be cost-effective to screen and verify the small set, rather than doing many runs to get single chromosomes. (On the other hand, if it's easy to split and reroute droplets, it may only be around twice as costly to obtain single number chromosomes compared to pairs of numbered chromosomes; you just take a guess and then do a confirmation test.)

- Sequencing / genotyping a clonal cell line.
  - (See the subsection "Sequencing a cell culture".)
  - Suppose you have a cell culture with several non-identical cell populations. For example, after the fusion step of MMCT, you have some cells with the original genome, some aneuploid cells, and some cells that have spontaneously corrected their ploidy in a way that incorporates the transplanted chromosome. You can isolate single cells and grow them, obtaining several clonal lines.

[5] Babahosseini, Hesam, Darawalee Wangsa, Mani Pabba, Thomas Ried, Tom Misteli, and Don L DeVoe. "Deterministic Assembly of Chromosome Ensembles in a Programmable Membrane Trap Array." Biofabrication 13, no. 4 (2021): 10.1088/1758-5090/ac1258. https://doi.org/10.1088/1758-5090/ac1258.

  - By sequencing a sample of cells from a clonal (genetically near-identical) cell line, you learn what exact chromosomes are in this line. I.e., you perform homolog identification.
  - But, the homolog identification is setwise. You know the set of homologs exactly, so you could write out the whole genome; but you can't necessarily point to any specific chromosome and say its number or sequence.
  - This can be useful on its own; e.g. in MMCT, one of the cell lines might already have your target genome.
  - This can also be useful in combination with other methods, e.g. with a method for individual chromosome identification. (See the subsection "Combining homolog and number identification".)
- Complementary sequencing with gametogenesis.
  - (See the subsection "Meiotic complementary identification".)
  - In theory, one could apply an in vitro gametogenesis method to single cells isolated into separate microwells. Then, by sequencing three of the four resulting haploid cells, the genome of the remaining haploid is known[6].
  - This reveals the set of exact chromosomes, without revealing the chromosome number of any specific chromosome. Since the set is haploid, if you learn the number of one of those chromosomes, you then know its exact DNA sequence.

More clarifications

- A chromosome identification method doesn't have to produce extreme confidence. Non-conclusive evidence is helpful (e.g. see the subsection "Screening and confirmation tests"). However, see the subsection "Isolating-ensembling methods require high-confidence number identification".
- There's a kind of "functional" chromosome identification performed by methods that can bind to specific DNA sequences.
  - (See the subsection "Targeted DNA operations".)
  - For example, a CRISPR system might perform some operation on some specified genetic locus. Targeted binding *can* perform regular chromosome identification—e.g. you could use a CRISPR system with a fluorescent molecule attached, and then you could directly see which cells contain chromosomes with the specified sequence, and even see where that specific chromosome is.
  - But further, targeted binding can perform some operation that doesn't inform you about any chromosomes. For example, you could cleave chromosomes that have a specified sequence. This would kill cells with those homologs, but wouldn't directly inform you about any specific chromosomes. (Ok, fine, technically it *does* inform you, by knowably increasing the relative prevalence of cells without the homolog; but unlike simple chromosome identification methods, this method intrinsically combines the *targeting* step with the step that applies some action targeted by the identification.)
  - We don't need to settle on an exact definition of chromosome identification, and we don't have to decide whether "functional" identification counts. It suffices to keep in mind that the purpose of chromosome identification is *Targeting*, towards specific chromosomes, the *Transmission* into and *Exclusion* from the final cell lines.
- There may be minor complications / irregularities regarding the sex chromosomes X and Y. For example, in some cases (e.g. with a male diploid), chromosome number identification would imply homolog identification, as there's only one X and one Y. However, in this article, we'll ignore this distinction, and deal with chromosomes as though they are all autosomes. We do this because the power of chromosome selection is nearly unaffected by ignoring selection on the sex chromosome, and because methods for identification should work fine for identifying the sex chromosome.

Context for chromosome identification

Why chromosome identification is hard

In general, working with chromosomes is hard.

Chromosomes are tiny (a couple microns by a few tenths of a micron). They aren't studied as much as cells because cells are much easier to work with. Cells want to intake and expel many small molecules, but not

[6] Benson-Tilsen, Tsvi. 'Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins'. Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins, 29 June 2022. https://tsvibt.blogspot.com/2022/06/non-destructively-sequencing-gametes-by.html.

large molecules like chromosomes, so transmission is difficult. Chromosomes are physically relatively fragile. Their epigenomic state is fragile. Measurement is usually destructive.

What are chromosomes?

Human chromosomes are 50 million to 250 million base pairs of DNA, wrapped around nucleosomes (histone octamers), and bound by many proteins.

Chromosomes that we handle in a chromosome selection protocol could come from several sources:

- Ordinary stem cell cultures.
- Special cell cultures, such as cultured embryonic stem cells, actual early embryos grown in vitro, or haploid cell lines (see the subsection "Haploid culture").
- Gametes, i.e. sperm or oocytes. Also, more generally, cells that have previously undergone meiosis.

For purposes of chromosome selection, one key question to ask about some given chromosomes is "Are these chromosomes in a gamete-like or zygote-competent epigenomic state?". See that linked section for an explanation of why this is a crucial question.

If we don't have a method to correct the epigenomic state of cells, then we cannot safely make a healthy baby from some cell unless that cell already has a reproductively competent epigenomic state. Therefore, a chromosome identification method has to either rely on an epigenomic correction method, or else it must preserve the epigenomic state of the chromosomes that it identifies.

Speaking very coarsely, a chromosome can be in these *epigenomic* states:

- Epigenomically correct.
  - That is, it has the genome-wide deactivation as well as the sex-specific imprinting (methylations, histone modifications) that matches what that chromosome should have if it were part of natural gamete DNA.
  - These chromosomes can be used to compose the genome of a zygote to potentially bring to term.
- Nearly epigenomically correct.
  - That is, it's almost in a correct state, but maybe has been recently demethylated, and/or has lost a few imprints.
  - These chromosomes might be usable for a live birth, given more ability to correct their epigenome and more understanding of what a correct epigenome has to be.
- Not epigenomically correct.
  - These chromosomes cannot be used for a live birth, without a method to correct their genome, such as in vitro gametogenesis.

Again speaking coarsely, a chromosome can be in these *structural* states:

- Metaphase.
  - A metaphase chromosome is tightly packed. It is a bit physically smaller than it was in interphase in terms of displacement volume, and it is geometrically much more compact, i.e. it is arranged as a blob rather than spaghetti.
  - A metaphase chromosome has largely inaccessible DNA. The DNA, and probably most of the nucleosomes, are covered by other parts of the chromosome (other regions of the DNA and/or other structural proteins). So, they can't be easily targeted with DNA binders without first being treated in order to loosen the protein structure.
  - Metaphase chromosomes are not connected to each other, but the copied sister chromatids are connected to each other. (They separate at anaphase, which might be the appropriate time to retrieve them.)
  - A metaphase chromosome is probably fairly strong, mechanically. It can withstand probably at least a piconewton, maybe much more (even up to 100 piconewtons or more) of force.
  - Metaphase chromosomes might require certain proteins and/or ATP to maintain their structure[7]; I'm unsure.
- Interphase.

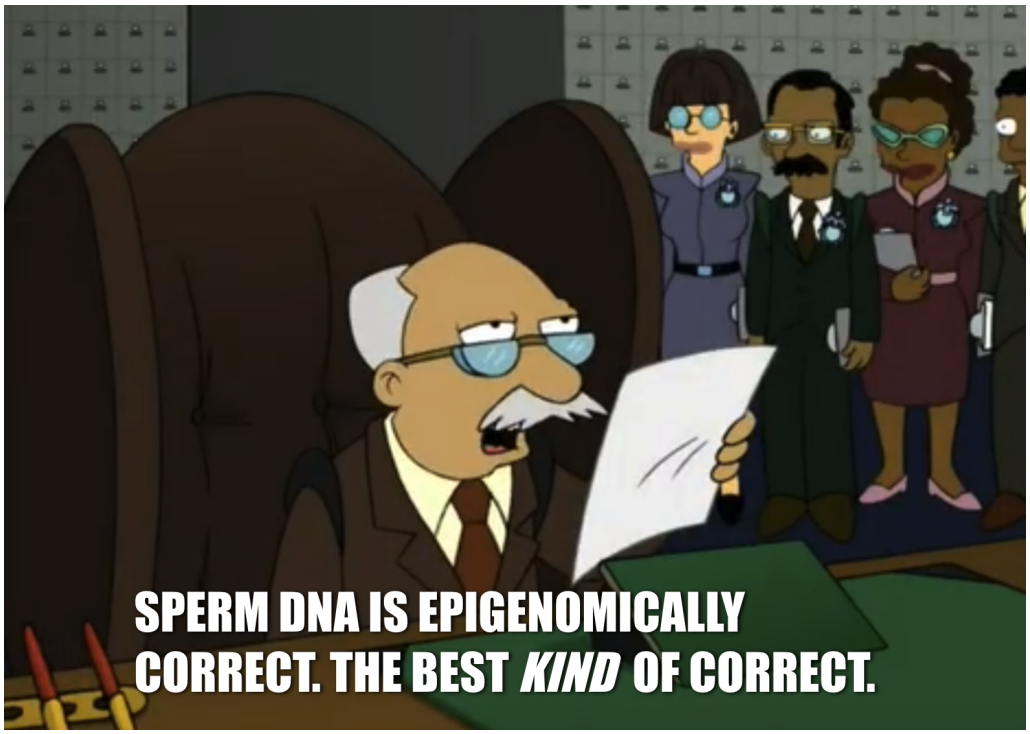
[7] Paulson, James R., Damien F. Hudson, Fernanda Cisneros-Soberanis, and William C. Earnshaw. "Mitotic Chromosomes." Seminars in Cell & Developmental Biology, Mitosis, vol. 117 (September 2021): 7–29. https://doi.org/10.1016/j.semcdb.2021.03.014.

  - Interphase chromosomes are basically the opposite of metaphase chromosomes, in the respects listed in the previous bullet point. They are not compact, they are mostly very accessible, they are not connected to a duplicate, and they are mechanically fragile but are maintained by the cell's repair mechanisms.
  - In particular, interphase chromosomes are probably only usable if they're inside live cells. Otherwise, they would be difficult to mechanically manipulate and easy to damage.
  - Many targeting methods (see the subsection "Targeted DNA operations") are probably generally easier to use on interphase chromosomes, since they are more accessible.

- Spermatozoon or immature spermatid.
  - Spermatid chromosomes may be easier to manipulate, as they are still disconnected; but they may have an incorrect epigenomic state.
  - Spermatozoon chromosomes are in the correct paternal epigenomic state. However, since they are protaminated, they are likely all stuck together. This is an obstacle for isolating these chromosomes, and for any methods that involve binding anything to DNA (or histones). The sperm nucleus would have to be loosened and pulled apart.

- Mature oocyte, pre-fertilization. I don't know about these. (Doing chromosome selection on oocytes is unlikely to make practical sense, at least until in vitro oogenesis is solved. See the brief discussion in "RNA-based genotyping for oocytes?".)

- Haploid culture or embryonic stem cell culture.
  - These are interesting in that they may have a correct-enough epigenomic state, while also passing through a metaphase state. So they might be more easily mechanically manipulated.
  - They could also be sequenced (see the subsection "Haploid culture") for setwise homolog identification.



We'll generally focus on metaphase chromosomes. That's because metaphase chromosomes are the most likely to be relevant to strong germline engineering: Chromosome selection on metaphase chromosomes can combine with other genomic vectoring methods such as iterated recombinant selection and iterated stem cell editing. Sperm chromosomes are also important because they might be used to bypass the problem of epigenomic correction, but that idea has additional obstacles (understanding epigenomic correctness and dealing with condensed sperm chromatin).

Cell culturing vs. isolating-ensembling methods

Generally, we can operate on chromosomes while they are in live cells or after they have been extracted from cells. This induces some tradeoffs:

- Epigenomic correctness vs. culturing.

  - There is a tradeoff between culturing and epigenomic correctness. The more you culture some cells, the more loss of imprinting occurs, and the more genome-wide reprogramming occurs.
- DNA damage.
  - Live cells provide many repair mechanisms, keeping chromosomes intact. Outside of cells, chromosomes are at risk of accumulating damage. Further, in a live culture, you can select for desired chromosomes.
  - On the other hand, cells slowly but continually accumulate DNA damage as they grow, which might be infeasible to correct.
- Enabling operations on chromosomes.
  - Live cells make several important operations on chromosomes easier.
    * Live cells provide a way to accurately duplicate full chromosomes, which is not otherwise feasible. That's an easy way to get more of the same chromosome to use for more attempts at some manipulation. In particular, this enables selection: You preferentially grow cells with desired chromosomes, or kill cells without those chromosomes.
    * Chromosomes in live cells pass through interphase, during which they are easy to target.
    * Live cells can sometimes aid *transmission* of chromosomes to other cells. For example, see cell electrofusion and MMCT.
  - For these reasons, live cells offer a lot of surface area for methods we could try for manipulating what sets of chromosomes are present in our cultured cell.
  - On the other hand, there are some key constraints in the live cell context. Besides genomic and epigenomic disruption, there may be a speed limit on *exclusion*. If you naively cut up a bunch of chromosomes in some cells all at once, in order to leave behind just your desired chromosomes, you're likely to kill the cell (and/or select for cancer) by activating apoptotic DNA damage responses. If you exclude chromosomes one at a time, it may take many weeks or months to exclude, say, 23 or 46 chromosomes from the cell. Those cells would be likely to accumulate a lot of DNA damage.

Because of these tradeoffs, it's not clear whether to do chromosome selection in a cell-culturing context. On balance, my inclination is towards *isolating-ensembling* methods. Those are methods where chromosomes are extracted from cells, identified, and then individually selected to be ensembled together in a new cell.

So, this article is written mostly having in mind chromosomes that have been extracted from cells.

Direct and indirect chromosome identification

Direct chromosome identification methods work by directly interacting with the specific DNA molecule to be identified. Indirect methods work by instead interacting with other molecules that are informationally entangled with the specific DNA molecule in question.

Both sorts of methods could work as part of a chromosome selection method. In fact, because many direct identification methods are destructive (e.g. sequencing), indirect identification methods may be the best available option in the context of chromosome selection.

- Examples of direct identification: sequencing, SNP arrays, staining, electrophoresis, Raman scattering. (See the section "Direct chromosome identification methods".)
- Examples of indirect identification: single-cell RNA sequencing, clonal sequencing, complementary identification. (See the section "Indirect chromosome identification methods".)

To a first approximation, every identification method is either direct or indirect. Suppose that *every* direct identification method available is either destructive, or else it cannot produce sufficiently confident calls about chromosomes. We need confident calls; see the subsection "Isolating-ensembling methods require high-confidence number identification". In this case, we have to use indirect identification with a destructive, high-confidence identification method.

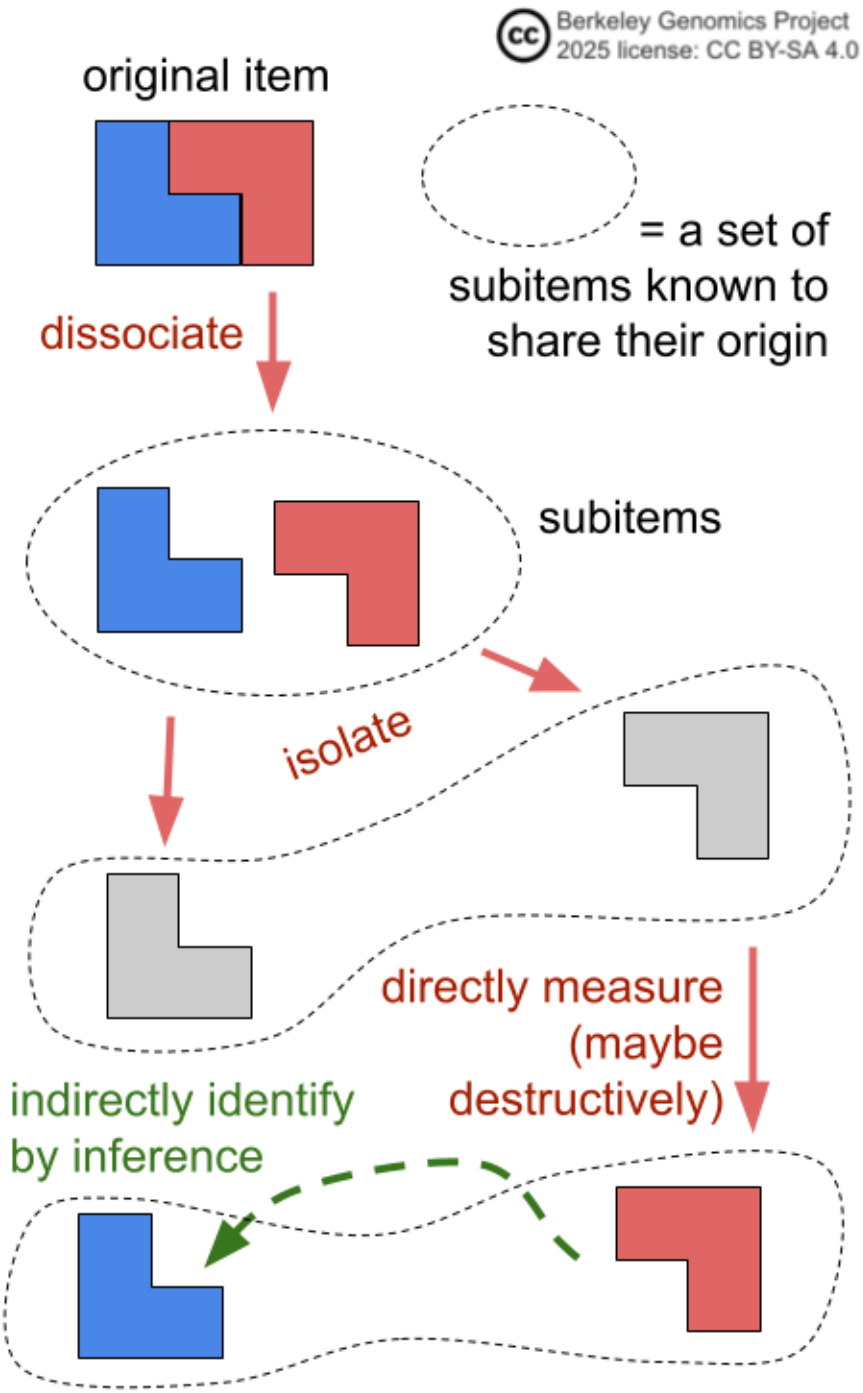
(Since "direct"/"indirect" is a bit conceptually incoherent, a more precise terminology might be "more/less direct", or "proximal/distal". Practically speaking, a direct chromosome identification method is one where the target chromosomes—the ones that will eventually end up in the final output cell—are physically operated on by measurement devices. An indirect identification method is one where the measurement devices instead physically operate on other molecules that have been spatially separated away from the DNA molecule in question.)

Indirect chromosome identification

Indirect methods work as in the following general schema:

1. There's some item.
2. The item is dissociated into two subitems through some known process.
    - One of the subitems is preserved.
    - The other subitem is measured (maybe destructively).
3. Something about the preserved subitem is inferred, using the fact of its common origin with the measured subitem, and using the results of the measurement.

A schematic illustration of indirect identification:



In general, indirect identification tends to be possible in both directions, assuming you're able to non-destructively separate the two subitems. You could swap which subitem is preserved and which is measured, and then still just as well make an inference from the measurement to something about the preserved subitem.

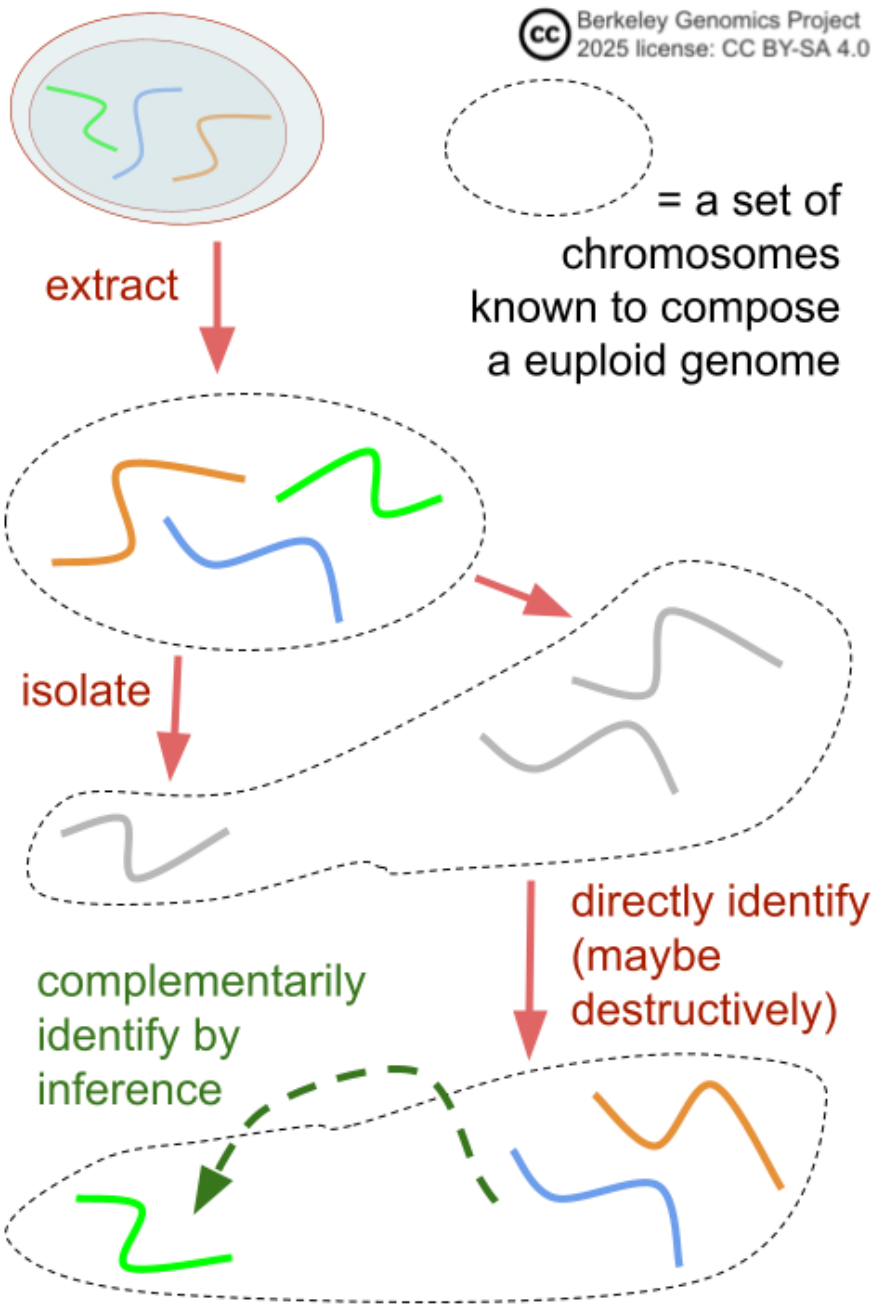
Examples of indirect identification:

- *Complementary chromosome identification.*

  - (See the subsections "Destructive identification and complementary identification" and "Chromosome-wise complementary identification".)
  - In chromosome-wise complementarity, the original item is a single euploid cell, and the subitems compose a partition of that cell's chromosomes. Since the two sets come from a single euploid cell, the set of numbers in one set tells us the set of numbers in the other set.
  - In meiotic complementarity, the original item is a diploid cell. The preserved subitem is a single one of the meiotic daughter haploids; the measured (destructively genotyped) item is the set of the other three daughter haploids. Since the two sets of haploid genomes come from the original diploid cell via meiosis, we know their total collective DNA content, and can make the desired inference. (See the subsection "Meiotic complementary identification".)
- *Sequencing / genotyping a clonal cell line.*
  - (See the subsection "Sequencing a cell culture".)
  - The original item is a cell culture that came from a single cell, grown through mitosis. The subitems are (1) a sample of cells from that population, and (2) the remainder of that population. Their common origin, through mitosis, from a single cell, implies that they have nearly identical genomes, so homologs can be identified this way.
- *Single-cell RNA sequencing for homolog identification.*
  - (See the subsections "Sequencing post-meiotic RNA" and "Combining homolog and number identification".)
  - Here, the original item is a single cell. The preserved subitem is the set of chromosomes from the cell. The measured subitem is the cytoplasm / nucleoplasm of the cell, dissociated from the chromosomes by membrane lysis and some sort of wash or chromosome manipulation, and then destructively RNA sequenced. Since the RNA in the plasm comes from the cell, sequencing the RNA tells us what homologs are in the cell's own genome (the preserved subitem).

Destructive identification and complementary identification

Direct identification methods tend to be destructive. For example, SNP arrays and DNA sequencing require chopping up the DNA, and staining tends to require damaging DNA in order to work. For the purpose of doing chromosome selection, it's useless to identify a chromosome if you also destroy it. To address this problem, a key class of indirect identification methods is *complementary identification* methods. An illustration of complementary identification:



In general, these methods work as follows:

1. Isolate a single source of chromosomes.
2. From that source, derive two or more sets of chromosomes, so that the union of those sets has a known genome.
3. Use a direct method to directly identify all chromosomes in all but one of those sets.
4. By complementarity, infer the contents of the remaining set of chromosomes.

A complementary identification method takes a (maybe *destructive*) direct identification method, and makes a *non-destructive* indirect identification method. This benefit comes at the cost of adding in more complexity, compared to a direct method. The direct identification method has to be applied more times (e.g. $23 - 1 = 22$ times or $46 - 1 = 45$ times, for chromosomes—though usually inexpensively batchable), and a more complicated apparatus and/or protocol is needed to separate and identify the complementary chromosomes.

- Example 1: chromosome-wise complementarity.
  - (See the subsection "Chromosome-wise complementary identification".)
  - Example protocol:
    - Step 1: start with a cell nucleus known to be haploid, e.g. the nucleus of a sperm.
    - Step 2: isolate one chromosome, with the remaining haploid minus one chromosomes in the other set.
    - Step 3: G-band the remaining haploid minus one chromosomes to identify their numbers.
    - Step 4: the number of the isolated chromosome is the number that doesn't appear in the karyotype from step 3.

  - This can also be applied to a diploid starting cell. In this case, the missing chromosome number appears once in the karyotype while all the other chromosome numbers appear twice.
  - This can also be applied setwise (see the subsection "Setwise identification"). Instead of isolating a single chromosome, you can isolate some set (perhaps a pair) of chromosomes, and G-band the remainder.
  - As mentioned in "Indirect chromosome identification", indirect identification generally works in both directions. This is true for chromosome-wise complementarity, as a setwise identification method: if you can reliably isolate a specific set A of chromosome numbers, then almost certainly you can also reliably isolate the complement of A, by just taking the remaining chromosomes (unless for some reason you destroyed them in the process of isolating A). Since complementary numbering works both ways, you can therefore get both A and its complement knowingly. If you ensemble together a set of chromosomes A from one starting cell line and the complement of A from another cell line, you've made a euploid genome from chromosomes from multiple cells, i.e. you've accomplished a form of partial chromosome selection.
- Example 2: meiotic complementarity.
  - Step 1: start with an isolated terminal spermatocyte, about to undergo meiosis.
  - Step 2: derive 4 meiotic child cells, i.e. 4 sperm; isolate one of them.
  - Step 3: quantitatively genotype the remaining 3 sperm, e.g. using a SNP array.
  - Step 4: infer the approximate genotype of the isolated sperm, according to which DNA segments only appeared once in the set of 3 sperm.
  - (See the subsection "Meiotic complementary identification".)

Indirect identification requires origin indexing

As described in the subsection "Indirect chromosome identification", indirect identification works by inferring (step 3) from the measurement of one subitem to something about the other preserved subitem. Crucially, to make this inference, we have to know the specific two subitems, as a set; we have to connect them in our knowledge as a pair.

If we don't know that these two subitems are a pair, we can't learn about the preserved subitem.

For example: Consider scRNAseq for homolog identification. (See the subsection "Sequencing post-meiotic RNA".) Imagine that you successfully perform single-well RNA sequencing for 256 wells, so you learn the homolog sets for each of the 256 isolated original cells. Separately, you use complementary number identification on each individual genome, so you have 256 chromosomes, each with known number. Do you furthermore in fact have 256 isolated chromosomes of known homology? No, not necessarily. To use the homology information, you would have to connect each individual RNA sequencing readout to each isolated chromosome. If you didn't preserve that information, you can't learn which homologs you have.

Thus: Indirect identification requires origin indexing. The preserved subitem and measured subitem have to be mapped to their origin (or at least, to each other).

Bulk methods tend to destroy indexing information. In other words, if you have a big blob of items from different origins, mixed together so their origins are indistinguishable, then you can't identify pairs of same-origin items. For example:

- If you have a large sample of naked chromosomes, you can't tell which sets of chromosomes originated from single nuclei. So you can't use indirect identification to number these chromosomes.
  - (The especially cheeky reader will note that if you isolate one chromosome from a bulk sample comprising every single chromosome from a sample of 100% euploid cells, and then sequence every single other chromosome, you can in theory indirectly identify the isolated chromosome. But, as they say, "In biology, cheekiness is rarely rewarded.". (Who is "they"? Me. Me is they. I am they. Them. Whatever. Stop.))
- If you have a sperm sample with many sperm, you can't tell which sets of four sperm came from a single primary spermatocyte. So you can't identify homologs by complementary gamete sequencing.
  - (Again, being cheeky, we can in theory imagine taking a sperm sample and then sequencing a large fraction of the sperm. We'll identify many cousin triplets—sets of three out of four sperm that must have come from a single meiosis event. Then we can start sequencing all-but-one sets of chromosomes from sperm. When we find an all-but-one homolog set corresponding to the fourth chromosome of a known triplet, we know the remaining homolog of our single isolated chromosome.)

There are a few ways to do origin indexing:

- *Isolate and process separately.*
  - In these methods, we isolate each single original item before we dissociate it into two subitems. We keep the subitems isolated from other subitems as we process them; after we've measured one subitem, we still know where the co-originating subitem is.
  - Example: clonal cell line sequencing. The original single-cell passage is the isolation step, and we grow the sample in its own area (e.g. a microwell).
  - Example: isolated single-cell microfluidics circuit. We isolate a single cell, lyse it, and then process the extracted chromosomes in a microfluidics circuit that's physically separate from, say, other circuits for other cells.
- *Barcoding / labeling items.*
  - In these methods, we first isolate the original items; then we label the items so that each distinct original item gets a different label. Only then do we dissociate the items into subitems. We can then process the dissociated subitems without keeping them separate. After we've made measurements, we can correlate a preserved subitem with the measurement taken from another subitem that has the same label. Since we know they share an origin, the inference for indirect identification goes through.
    - These methods have the benefit of letting us do operations after labeling in bulk, or without physical separation. This is less costly, requiring fewer copies of apparatuses (which would be needed to process single isolated items in parallel).
    - On the other hand, we have to do some additional operations to label the items and later read and correlate the labels; and some labels are destructive. That adds complexity to the protocol.
  - Example: DNA barcoding. In some contexts, we can integrate DNA markers into a cell's genome in a way that is different for different cells. Then, after we dissociate subitems from such cells (e.g. through meiosis or chromosome extraction), we can sequence many subitems in bulk. By comparing barcodes, we could identify chromosomes in subitems that we haven't yet sequenced[8]. This has the great advantage of sequencing/measuring in bulk; but it has the great disadvantages that we're modifying the DNA, and that we have to use FISH or similar to locate the subitem with the barcode corresponding to known chromosomes.
  - Example: droplet dyeing. Microfluidic circuits tend to run at high throughputs—hundreds or thousands of droplets per second. To take advantage of that, we'd want to put many droplets through each circuit. It might be possible to inject a set of dyes or other labels into droplets, uniquely according to their cell of origin. Then we can process droplets from many cells with a single circuit, and later locate droplets with identified chromosomes.

Combining multiple identification methods

In many cases, it could be helpful to use multiple different chromosome identification methods in a single protocol. The following subsubsections list sorts of reasons that multiple identification methods might be helpful or even practically necessary.

In fact, this is likely the situation, given the reasons listed below. So, we can't be deadset on finding a single identification method that meets all our needs. This unfortunately cuts against the constraint that chromosome selection methods should be efficient to develop in a research project: it doesn't follow the strategy of keeping the number of heterogeneous elements (steps, apparatuses, skillsets) small.

Combining homolog and number identification

One key use for multiple identification methods is to fully identify a chromosome (both number and homolog), by inferring the number and homolog through different lines of evidence. More precisely, we apply a protocol matching this schema:

1. Obtain an isolated haploid nucleus, e.g. via meiosis.
2. Apply two methods:
   a. Apply a nondestructive setwise homolog identification method to (approximately) learn the haploid genome. We learn the set of homologs in the haploid genome, but we don't learn about any

[8] Benson-Tilsen, Tsvi. 'Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins'. Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins, 29 June 2022. https://tsvibt.blogspot.com/2022/06/non-destructively-sequencing-gametes-by.html.

one specific chromosome.
   b. Isolate one chromosome and apply a nondestructive chromosome numbering method to learn its number.
3. Infer most of the exact DNA sequence of the isolated chromosome from 2a and 2b.

Steps 2a and 2b might be performable in parallel or in either order. A similar schema would work for diploid cells, but would require identifying the homolog of the non-isolated chromosome (or doing setwise identification to get a pair of chromosomes of the same number). (For examples of methods for 2a and 2b, see the subsections "Sequencing post-meiotic RNA" and "Raman spectroscopy".)

Screening and confirmation tests

Another key sort of combination is a screening test followed by a confirmation test. We use the screening test to direct our attention to candidate chromosomes that are likely to be the target chromosome; then we use the confirmation test on those candidates to reach high confidence that we've correctly identified them.

As described in the subsection "Isolating-ensembling methods require high-confidence number identification" below, if we're going to ensemble a euploid genome with non-negligible probability, without culturing, then we have to be highly confident in our chromosome numbering guesses. So at some point we have to use an identification method that can reach high confidence. For example, complementary identification using G-banding (or ordinary DNA sequencing) on the non-target chromosomes should work. However, while G-banding is a reasonable method, the full procedure has non-trivial costs in time (a day, perhaps) and labor.

In order to avoid paying the cost for high-confidence identification many times over, we want to apply the high-confidence method as few times as possible. So instead of applying it to every candidate chromosome, we first narrow our search to chromosomes that pass a cheaper, quicker, lower-confidence test. For example, we can simply inspect a chromosome in an ordinary light microscope. This is cheap and fast, and might be able to coarsely distinguish the longest chromosomes from medium-length or at least the shortest chromosomes.

Examples of possible cheap screening test methods: mass sorting, visible size, dielectrophoresis, Raman scattering.

Examples of possible high-confidence confirmation test methods: G-banding, SNP arrays, sequencing, fluorescent markers.

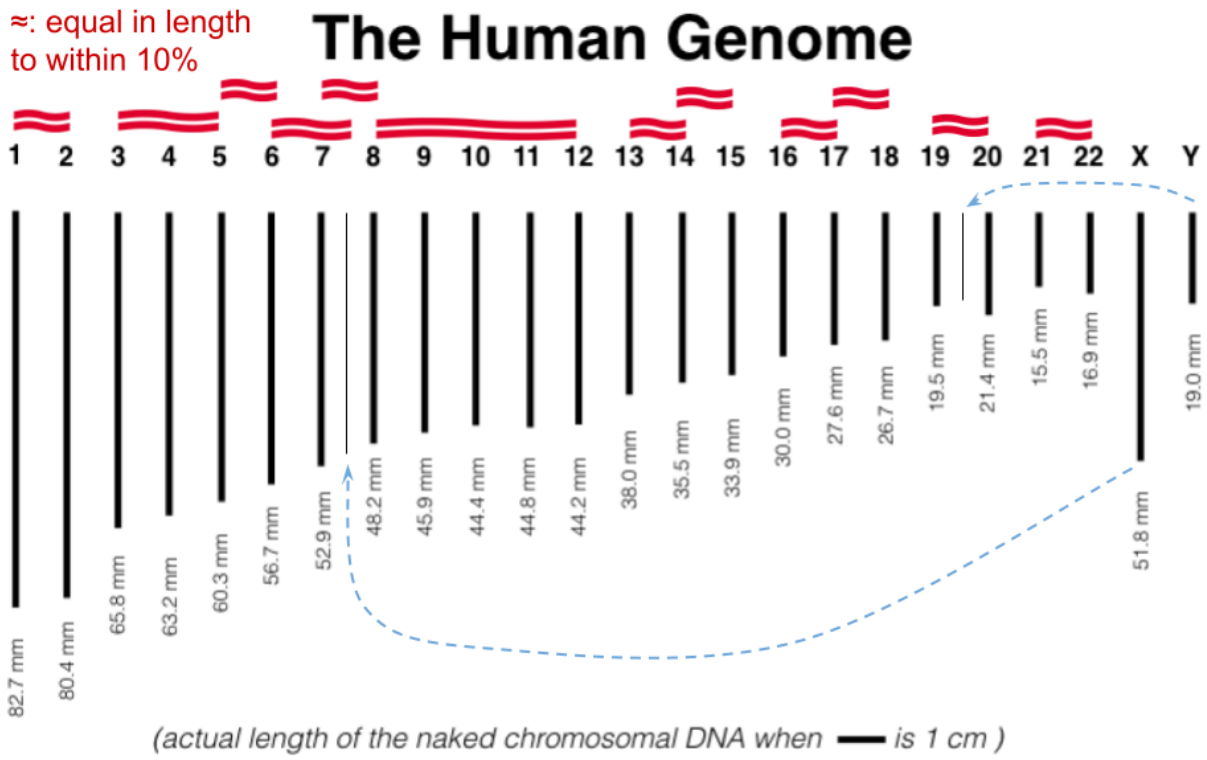
In the same vein as the split screening / confirmation strategy, and the split homolog / number strategy: Suppose we have a diploid cell that we can culture. Through ordinary clonal sequencing, we know the cell's full genome. Now we want to isolate a single known chromosome. We might produce many isolated chromosomes and then identify their numbers with a fast and inexpensive method, e.g. complementary staining, which tells us number but not homolog. Once we've located a few isolated chromosomes with known number, we could use a more expensive method on the complementary set of chromosomes to identify the specific homolog that we've isolated. E.g. we could use SNP array genotyping to find which homolog is in the complementary set (and therefore which homolog we have isolated).

Jointly discriminative feature sets

Some chromosome identification methods are fully discriminative on their own. SNP array genotyping would be able to (destructively) distinguish, with very high confidence, any two chromosomes of different number (or even any two homologs with large segment differences).

On the other hand, some identification methods might be generally good (inexpensive, fast, non-destructive), but fail to fully distinguish all chromosomes, even by number. It could be that by combining two or more of those methods, we can distinguish all the chromosomes by their joint measurements from all the methods. Even though no single method is fully discriminative on its own, the set of methods together could be jointly fully discriminative.

For example, it could be that one could use simple light microscopy to distinguish chromosomes by apparent length; but only within 10%. Human chromosomes vary greatly in length, but there are several sets of chromosomes that are within 10% of each other in length:



(Modified from figure 1.3 in Gallegos (2022)[9]. © publisher. Note: This is not quite the right picture to show: This picture shows the length of the DNA if it were stretched all the way out; but what we care about is the visible shape of a highly compacted metaphase chromosome. But the qualitative points will stand. A chromosome's physical length in general tends to be close to proportional to the number of base pairs in its DNA[10], and for example the physical length of chromosomes $\{10, 11, 12\}$ in metaphase are in fact quite close together[11].)

Each red "approximately equal" symbol ( ) stretches over two or more chromosome numbers, indicating that those chromosomes aren't distinguishable by length to within 10%. This diagram shows several pairs of chromosomes that aren't distinguishable; a few triplets; and even a quintuplet (chromosomes 8 through 12). Note especially chromosomes 10—12, which are very very close in length.

So, chromosome length, as visually measured, isn't fully discriminative (unless it's somehow extremely accurate). Suppose hypothetically that Raman spectroscopy produces spectra for chromosomes that also have significant ambiguity, meaning that Raman spectroscopy is also not fully discriminative. But suppose further that these ambiguities are *disjoint* from the ambiguities in length. In other words, suppose that chromosomes 8 through 12 have Raman spectra that are all pairwise distinguishable; and likewise for all the other sets of chromosomes that aren't noticeably different in length.

In this hypothetical, chromosome length together with Raman spectra composes a jointly discriminative feature set. Any two chromosomes of different numbers will have noticeably different lengths, noticeably different Raman spectra, or both.

An example that might be practically useful is doing number identification by complementary staining, and using multiple stains to avoid having to analyze banding patterns. (See the subsection "Flow cytometry".)

(In general, measuring multiple features in a single protocol is a standard approach. E.g. there is "multi-parameter analysis" in flow cytometry, which might measure different sorts of scattering (side vs. forward angle), use different laser light wavelengths, and might use both lasers and ligand fluorophores. There is also "multiomics" (even single-cell multiomics) for measuring the molecules inside cells, which might be of multiple types—DNA, epigenetic chemical or structural differences, RNA, or proteins.)

[9]Gallegos, Maria. Fantastic Genes and Where to Find Them. Updated 2022-09-13. Accessed 16 February 2025. https://bookdown.org/maria_gallegos/where-are-genes-2021/#preface.

[10]Kramer, Eric M., P. A. Tayjasanant, and Bethan Cordone. 'Scaling Laws for Mitotic Chromosomes'. Frontiers in Cell and Developmental Biology 9 (June 2021): 684278. https://doi.org/10.3389/fcell.2021.684278

[11]Van Dyke, D. L., Maria J. Worsham, L. J. Fisher, and L. Weiss. 'The Centromere Index and Relative Length of Human High-Resolution G-Banded Chromosomes'. Human Genetics 73, no. 2 (1986): 130–32. https://doi.org/10.1007/BF00291602

Ground truth tests

It may be that a complete chromosome selection protocol (e.g. a minimal proof of concept or an application-ready protocol) uses only a single chromosome identification method. But even in that case, we might want to use additional identification methods as a way of getting ground truth measurements.

This is similar to the combination of screening and confirmation tests described in the previous subsubsection, but in that case, both tests are part of the complete protocol, because the screening test cannot provide the needed level of confidence. Ground truth tests, on the other hand, aren't part of the complete protocol, or at least not part of the protocol that gets executed hundreds or thousands of times in a single run of chromosome selection. Rather, ground truth tests are used before the actual CS run (actually assembling a complete chromosome-selected genome), as a tool to set things up so that the actual CS run goes well.

In general, ground truth is used to get feedback about which setups work and which don't; that directs us about which adjustments to make, and tells us when we're done. Examples of uses for ground truth chromosome identification tests:

- *Feedback while we're developing the basics of a CS protocol.*
  - Generally, ground truth tests provide a target to aim at when we're varying the setup that we're testing. Some examples:
  - *Checking when a protocol works end-to-end or not.* You run the protocol using the in-protocol identification method; then you apply a separate ground truth identification method to the result. This way, you know which specific implementations of the protocol work well or poorly. (This is an example of the general constraint that chromosome selection methods have to be as measurable as possible; we want to know what's even happening in our apparatus.)
  - *Debugging a multistep protocol by confirming results stepwise.* Say you successfully isolate and identify single chromosomes, and now you additionally try to ensemble them together. But the end result is not what you were trying to get. So, you can (re)run a ground truth test on the isolated chromosomes, or on the final ensemble, or on other intermediates (e.g. partial ensembles of some but not all of the target chromosomes), to identify which is the first step where something goes wrong.
  - *Feedback for training faster identification methods.* For example, suppose we want to develop a Raman spectroscopy method to non-destructively identify chromosome number. We get spectral data, but there's no obvious way to identify the chromosome number based on that data. So, we build a training set by applying the Raman spectroscope, followed by a (perhaps destructive) ground truth identification method; then we can find mathematical functions from spectroscopy data to chromosome numbers that recover the ground truth. In the final CS protocol, we apply the trained Raman chromosome identifier, but we don't apply the destructive ground truth identification test (P.C. Hodgell notwithstanding).
- *Recalibrating a working CS protocol for new conditions.*
  - Even once we have a minimal working complete CS protocol, the conditions where we apply the protocol might change. When that happens, the identification method we were using might break a little bit. In many cases, the method is still basically sound, but only actually works if we tweak some parameters to compensate for the changes. In this sort of situation, we'd use a ground truth identification method to guide our adjustments.
  - For example, many adjustments might change the exact spectrum of light scattered by a given sample of molecules; e.g. droplet size, flow rates, buffer media, apparatus materials, and so on. Those parameters might have to be tweaked, which would change the direct measurements collected by a spectroscopy setup. The measurements might be a simple transformation of the measurements in the previous setup, e.g. slightly attenuated across the whole spectrum. In that case, we can collect a small amount of ground-truth data in the new setup, and figure out which simple transformation to apply to recover the previous predictive accuracy. (If the changes are not simple, and can't be corrected by an easy transformation, then the situation shades away from mere recalibration, into more basic development. We'd have to stop and collect much more ground truth data, in order to retrain a predictor from scratch. This is an example of violating the *developability* constraint on chromosome selection protocols.)
- *Verifying and demonstrating a working protocol.*
  - Say we have a complete CS protocol. It's ready for the next step; either we're going to start developing a further refinement, or we're going to actually apply the protocol to perform chromosome selection on cells for some use. But before we start an actual run of chromosome selection, we want to verify that the protocol works. So, we do a test run and then apply a ground truth test

to verify that the resulting ensemble of chromosomes has the numbers, and maybe homologs, we intended.
  - Further, we may want to demonstrate to other researchers that the protocol works; we'd do that with a ground truth test.

Isolating-ensembling methods require high-confidence number identification

Say we're trying to ensemble a euploid genome. We ensemble together several sets of chromosomes. These sets should form a partition of a euploid genome. If they don't, then the ensembled genome will not be euploid. If we make even a single incorrect call about the set of chromosome numbers in some isolated set of chromosomes, that would almost certainly disqualify the resulting ensemble (unless by a huge coincidence we made another incorrect call that exactly complemented the first incorrect call).

For isolating-ensembling methods, we don't get a chance to correct mistakes.

- That means we have to ensemble a whole euploid genome in one shot.
- To be more precise, this is not exactly right; only non-culture-based methods are required to get it right in one shot. In contrast, we could have an isolating-ensembling method that *also* uses cell culturing. In this case, we could ensemble a genome, produce and culture a cell line with that genome, and then sequence the genome to verify. Even if it's not our target genome, it is likely close, e.g. one chromosome is wrong. We can then do the whole procedure again, only trying to replace that one incorrect chromosome, which should be much easier.
- Another way this isn't exactly right is that, conceivably, we'd want to create many many ensembles, and then only later verify their genomes; this could be the case if we have a fast, reasonably accurate screening test, but only a pretty slow confirmation test (see the subsection "Screening and confirmation tests"). But even in that case, we could just wait for the verifications of individual chromosomes and then quickly ensemble the one correct genome.

Therefore, before we add a set of chromosomes to the ensemble, we have to be quite confident that we've identified the numbers of the chromosomes in that set.

Illustration of needed confidence

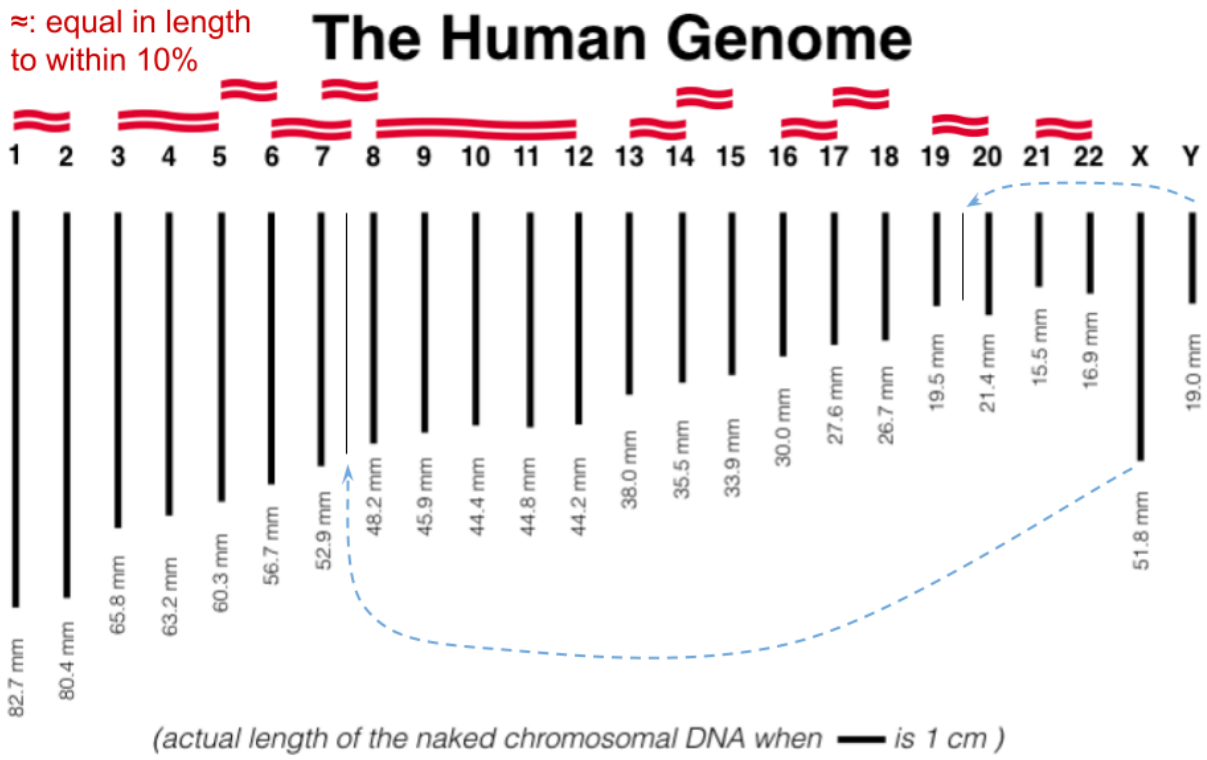
Suppose, for example, that we're making a haploid human genome of 23 chromosomes by adding single isolated chromosomes. Suppose we are 90% confident in each call of a chromosome number. What is the probability that we get a euploid genome? It is:

$$P(\text{euploid genome}) = P(\text{every call correct}) = P(\text{one call correct})^{23} = 0.9^{23} \approx 0.089$$

In other words, more than nine times out of ten, we ensemble an aneuploid genome. If we are using this haploid genome as a substitute for a male genome, and we have current-day numbers of oocytes available, this is an unacceptable ratio; we'd be left with at best a few, perhaps zero, euploid zygotes, and probably zero successful implantations.

On the other hand, if we're 97% confident in each call, then we have roughly a $0.97^{23} \approx 0.5$ chance of getting a euploid genome. That's probably acceptable, though higher would be better of course (for comparison, 99% confident calls gives roughly a 0.8 chance of a euploid ensembled genome).

Note that these numbers can be dominated by a few low-confidence calls. For example, suppose we make calls just based on visible length, with a few percentage points of error. (Like the following diagram, but somewhat more precise.)



(Modified from figure 1.3 in Gallegos (2022)[12]. © publisher)

Consider chromosomes 1 and 2. These chromosomes are within 3% of each other in length. That means if we go strictly off of length, and try to guess the number of a single isolated chromosome, we have about a 0.5 chance of getting it right. Now consider chromosomes 10, 11, and 12. Again, these are all within 3% of each other. So we'd have a 0.33 chance of getting it right. If we tried to get one of each from those two sets, we'd have about a $0.5 \times 0.33 \approx 0.17$ chance of getting both right; and an even lower chance of getting a full euploid genome.

If we only have that degree of accuracy available from our measurements, the correct strategy is not to make guesses about single isolated chromosomes. Instead, if possible, we'd want to make small sets of chromosomes. This would achieve setwise chromosome selection. If the groups can be kept small, the selection power of the resulting chromosome selection method is not decreased too much; see the brief discussion in "Setwise identification". If the groups can't be kept small, then we have a much smaller total number of groups, which does significantly decrease the selection power.

For this reason, isolating-ensembling CS methods usually require either

- quite confident calls about chromosome numbers, or else
- a small number of partitions, i.e. only a few sets of chromosomes are isolated, identified, and then ensembled.

Examples: bulk sorting by mass via centrifugation, visual length-based identification, and Raman spectroscopy each might be unable to make confident enough calls.

Setwise calls vs. uncertain calls

Bulk methods, i.e. methods that do not maintain origin-indexing information, are worse in terms of confidence.

To see why, first suppose that we are using a non-bulk method, and suppose that our direct chromosome identification method is low-confidence. More specifically, suppose that it leaves ambiguities, so that we can call sets of chromosomes with high confidence, but we can't resolve individual chromosomes within sets. E.g. suppose we can call chromosomes by length to within a 2% length variation. In this case, we can call all chromosomes except that we can't distinguish chromosomes $\{10, 11, 12\}$.

By assumption, we can index chromosomes to their cell of origin. Therefore we can do setwise chromosome identification and selection. In other words, we can select an input cell that has a high-scoring set of

[12]Gallegos, Maria. Fantastic Genes and Where to Find Them. Updated 2022-09-13. Accessed 16 February 2025. https://bookdown.org/maria_gallegos/where-are-genes-2021/#preface.

chromosomes $\{10, 11, 12\}$, and then take all three of those chromosomes from that cell to put into our final cell. Since we know the *exact set of chromosome numbers* with high confidence, the ensembled genome will be euploid with high confidence, as we needed. In this way, we can implement setwise CS, which is nearly as powerful as full CS (see the subsection "Setwise identification").

On the other hand, now suppose we are using a bulk method, i.e. we don't have indexing information. We can assume that we can identify a mass of chromosomes that is high purity, i.e. almost all chromosomes in this mass have the same length, up to a 2% tolerance. Again this gives us high-confidence numbers for most chromosomes. However, again we have ambiguity for chromosomes $\{10, 11, 12\}$.

In this case, the ambiguity is worse. Can we just take a set of three chromosomes, so that we'll get one each of $\{10, 11, 12\}$? No. The correlation between different samples from the mass of chromosomes would amount to origin-index information.

What happens if we take three chromosomes from this mass, at random? We get three independent identically distributed samples from a $\{1/3, 1/3, 1/3\}$ distribution over $\{10, 11, 12\}$. What's the probability we get a set of chromosomes with numbers $\{10, 11, 12\}$? Each ordering

$$(10, 11, 12); (10, 12, 11); (11, 10, 12); ...$$

has probability $1/3^3 = 1/27$ of occurring, and there are $3! = 6$ such orderings. So we have a $6/27 = 2/9 \approx 0.22$ chance of getting a euploid set $\{10, 11, 12\}$.

That is already a pretty bad rate of getting euploid genomes. Things get even worse quickly if the ambiguity sets are bigger, or if there are multiple ambiguity sets. In general, if we sample randomly k times from an ambiguity set of size k, there are $k!$ orderings of a euploid set and each one has probability $1/k^k$ of occurring. Here are these probabilities:

FACT(ROW()-3) / POWER(ROW()-3, ROW()-3)	
size of the ambiguous set	probability of sampling a euploid set
k	k! / k^k
1	1
2	0.5
3	0.222
4	0.094
5	0.038
6	0.015
7	0.006
8	0.002
9	0.001
10	0
11	0

So a plausible ambiguity set like $\{9, 10, 11, 12\}$ already upper bounds us below 10% for how likely our final cell is to be euploid. If we have two ambiguities then we have to get them both right. E.g. if we have ambiguity for $\{1, 2\}$ and for $\{10, 11, 12\}$, then we're upper bounded to have less than a $0.5 \times 0.22 = 0.11$ chance of sampling a euploid set.

Note that the meaning of "setwise identification" is a bit nebulous, because it's a heuristic guess about a maybe very complicated optimization problem. See the next subsection, "The full question of the power of an identification method". But my guess (in retrospect) is that these complications don't matter too much in practice.

The full question of the power of an identification method

There's an interesting, but probably largely irrelevant, math question here. The question is, given a chromosome identification method, how confidently can you assemble a euploid genome?

This is a complicated question.

You can't just model the measurement of a single chromosome. There may be important correlations between your beliefs about different chromosomes coming from a single cell. If you observe all the chromosomes coming from one euploid cell (e.g. by taking a light microscopy picture of a haploid cell's contents), what you end up with in general is a posterior joint probability distribution over all the numbers of the chromosomes; i.e. a distribution over all 23! assignments of numbers to chromosomes.

To add more complexity, there's a distribution over your possible post-measurement posterior distributions, induced by how the measurement method interacts with all possible cell contents / experimental conditions. You have to pick a policy for how to select one or more chromosomes from several cells, to ensemble into one cell.

So:

- Figuring out how good a given policy is for a given measurement, is a complicated question;
- figuring out what the *best* policy is for a given measurement, is an even more complicated question;
- and figuring out which measurement method has the *best* best policy, is a yet more complicated question.

Characterizing these distributions fully could be very complicated.

But in practice, I'm just using set size and confidence as heuristic summaries of distributions over posterior joint distributions over identifications. My guess is that in practice you just want to use an identification method that gives confident singleton calls for chromosome numbers, maybe in combination with euploid-genome-setwise calls for homologs.

The approximation I'm using is based on the heuristic that, even if we do use some setwise calls, we'll just use some fixed partition of chromosomes. From each cell, we extract a set of chromosomes which composes one of the sets in the partition; and we do so with high confidence, i.e. we're confident that we got exactly that set of chromosomes.

Confidence of intact DNA

As a final note, this subsubsection isn't strictly about chromosome identification, but it's worth stating here.

The same logic just discussed will also apply to DNA damage. It's acceptable for a CS method to have some chance of introducing harmful DNA damage to the genome, because we always will finally verify the genome of an embryo before implanting. However, if most embryos created with the CS method have unacceptable levels of DNA damage, the CS method is no good. (This is less of a constraint if we have IVG, because we can culture and verify, and then make gametes.)

If a protocol causes high rates of damage to chromosomes, then almost all embryos will have unacceptable damage. As a crude model, to illustrate, suppose that 90% of the time a chromosome is fine, and 10% of the time it is unacceptably damaged. (In real life damage is more quantitative, not binary.) Then the logic from the above discussion is the same: There's less than a 10% chance that you get a haploid genome without unacceptable damage.

Many methods run risks of causing DNA damage (or other damage, such as epigenomic disruption). Therefore, this constraint is actually felt: some methods will have to be ruled out because of the rate of damage they cause.

Note, though, that some damage is acceptable because it can be filtered out before ensembling chromosomes together and doing zygogenesis. For example, suppose we identify a chromosome as chromosome number 1, but then it breaks right in half, right in the middle. We should then be able to notice simply by light microscopy that we have two equally-sized short items, when we're supposed to have one large item. We'd have to try again to find an intact chromosome 1, but at least we wouldn't ensemble the broken chromosome together with our other chromosomes. Thus we avoid destroying our other work identifying chromosomes, and we don't waste a donor egg.

Tables of methods

This section lists criteria by which to judge chromosome identification methods, and then gives summary evaluations of methods.

Criteria for identification methods

Most good to least good: Green Yellow Orange Red

- *Refined.*
  How well-developed / standard / common / widely-known is the protocol?
  - Green: practiced by many scientists for many years with known parameters, and applied to chromosomes.
  - Yellow: somewhat more niche; or widely practiced but not applied to chromosomes.
  - Red: hypothetical, or presented in one or two papers with no replication or followup.
- *Price.*
  How costly is the method?
  - Specifically, how much does it cost to identify a single chromosome or a single batch of chromosomes (e.g. 22 of 23 chromosomes from a cell)?
  - This is a very unconfident attempt to guess the *unit* price of identification, *after* developing the protocol. So, it's hypothetical.
  - These are very imprecise and unconfident guesses. Also, they may have a vague meaning (especially for bulk methods).
  - Green < $10 < Yellow < $100 < Orange < $1,000 < Red
- *Time.*
  How long does it take to get answers about one round?
  - Green: minutes or less
  - Yellow: many minutes, up to several hours
  - Orange: a day or two, maybe
  - Red: longer than a day or two
- *Identifies numbers.*
  The method can identify the number of chromosomes.
  - Green: Yes, Red: No
  - (See the subsection "Number vs. homolog identification".)
- *Identifies homologs.*
  The method can identify chromosomes by alleles / crossovers.
  - That is, the method identifies which of multiple homologous chromosomes you have; i.e., it can distinguish chromosomes of the same number based on allele differences.
  - Green: Yes, Red: No
  - (See the subsection "Number vs. homolog identification".)
- *Set size.*
  How big of a set gets identified?
  - E.g. do we say "this one chromosome is such and such number", or do we say "this haploid cell contains chromosomes with these crossovers, but IDK which chromosome is which".
  - The actual meaning of this is sort of undefined. (See the subsections "Setwise identification", "Setwise calls vs. uncertain calls", and "The full question of the power of an identification method".)
  - Green: This identifies all single chromosomes.
  - Yellow: This identifies most single chromosomes, and some small sets.
  - Orange: Several small sets or some midsize sets.
  - Red: This only identifies cell-sized sets of chromosomes. (Note that for homolog identification, red is fine, especially if it's haploid; setwise homolog identification plus single chromosome number identification equals individual homolog identification.)
- *Confident.*
  How confident are the identification calls about chromosomes?
  - More specifically, what is the feasible *ceiling* of confidence? For some methods there could be tradeoffs, where you can wait longer or spend more resources to get higher confidence calls. Those tradeoffs are ignored here.
  - (See the subsection "Isolating-ensembling methods require high-confidence number identification".)

  - Green > 99% > Yellow > 95% > Orange > ~80% > Red
- *Saves DNA.*
  Does the method keep the chromosomal DNA intact?
  - Green: Yes
  - Yellow: Yes, but maybe not / maybe there are some mutations / maybe there's some telomere damage
  - Orange: No, the DNA of the chromosomes is likely significantly damaged, e.g. by accumulation of many mutations or oncogenic mutations, or by one or more double-strand breaks
  - Red: No, the DNA of the chromosomes is always totally destroyed (broken into many pieces)
- *Saves epigenetic state.*
  Does the method keep the chromosome in a gamete-like or zygote-competent epigenetic state?
  - Note that if human in vitro gametogenesis becomes feasible, we don't need to worry about this. Hopefully this will happen at least in the long-run if not sooner.
  - Green: Yes, the broad genome-wide epigenetic state is preserved, and also the sex-specific imprinting is exactly preserved
  - Yellow: Yes, but maybe not / maybe some imprinting is lost
  - Orange: Unknown / significant chance of significant epigenomic disruption
  - Red: No, the state is totally disrupted; the resulting cell is no better than, say, any normal pluripotent stem cell
- *Saves structural proteins.*
  Does the method keep the structural proteins integrated with the chromosome?
  - Some methods might loosen, digest, or strip away proteins that are integral to the mechanical structure of the chromosome. For example, metaphase chromosomes have DNA loops created and maintained by condensins. Without these proteins in their natural state, a chromosome might be more exposed to be damaged when it's handled as part of a chromosome selection protocol. See Paulson et al. (2021)[13]. They state: "Possibly, the chromosome as a whole becomes more susceptible to shear when connections within the chromosome scaffold (see below) are weakened.".
  - (An additional problem with stripping away histones specifically, is that doing so might render a chromosome impossible to epigenetically correct. Histones seem to plausibly carry some important imprints; if the histones are gone, they can't be modified.)
  - Green: Yes, it leaves them all where they are (assuming the rest of the chromosome selection protocol does so)
  - Yellow: It may loosen or disrupt some proteins, possibly leaving the chromosome more vulnerable to shearing forces
  - Red: The protocol probably removes large portions of proteins including core structural proteins such as histones

- *Direct/indirect.*
  Is it a direct or indirect identification method?
  - (See the subsection "Direct and indirect chromosome identification".)

Table of most methods

Below is a table listing most of the chromosome identification methods discussed in this article. Each method is judged by each criterion listed in the previous section. The methods excluded from the table seem to be probably not so important for chromosome identification.

Minor notes:

- Most of the judgements shown are imprecise guesses, sometimes blurring between different specific methods. The hope is just to give a partial impressionistic overview; for more detailed information, see the citations in the linked sections.
- All direct methods can be done without cell culturing (ignoring prior or posterior steps). Of the indirect methods, haploid and diploid culturing of course require culturing, while targeted elimination and meiotic complementation are border cases.
- Indirect methods involve using some unspecified direct identification method. For indirect methods, the evaluations given below are about what is *added* by the indirect method, ignoring the direct method.

[13] Paulson, James R., Damien F. Hudson, Fernanda Cisneros-Soberanis, and William C. Earnshaw. "Mitotic Chromosomes." Seminars in Cell & Developmental Biology, Mitosis, vol. 117 (September 2021): 7–29. https://doi.org/10.1016/j.semcdb.2021.03.014.

E.g. the evaluations show the cost of the indirect method, over and above whatever the cost of direct identification that's used.

- For indirect methods, criteria about disruption of the chromosome's state (DNA, proteins, epigenomic) refer to the chromosome that is *indirectly identified*, not to the chromosomes that are directly measured (and probably discarded).
- Rows from this table are reproduced in the corresponding sections.
- Most good to least good: Green Yellow Orange Red
- The row with column labels is repeated just for convenience.

Direct methods										
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Light microscopy	not much applied to chrs		seconds or minutes			ambiguous sets up to 4+				
Staining		labor	hours/ 1 day			some stains roughly give singles			some fixing methods disrupt histones	fixation methods loosen proteins
Fluorescen in situ hybridizat		?	a day?		with more design work	1			denatures DNA, probably removes many histones	
CRISPR-d systems	refined, but not much with chrs	design work?	minutes or hours?	with design work	with more design work	1		probably ok?	? mildly denatures DNA, probably removes some histones	? fixation methods loosen proteins
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Low-cover DNA sequencing		~$100?	a day or two?			1				
SNP array genotyping		~$100?	a few days?			1				
Centrifuga		??	minutes or hours?			ambiguous sets up to ~4 by mass	bulk method; not confident	high shear forces?	?	?
Electrokin	not much applied to full chrs	??	hours?	hasn't been shown		multiple features might ID singles?	?	shear forces?	?	?
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Hydrodyna methods	not much applied to chrs	??	minutes??	hasn't been shown		could be ok using mass and shape?	?	shear forces?	?	?
Raman spectrosco	one paper	?	minutes?? but low throughput	claimed		might be precise? claimed 1 vs. 2	might be confident?	laser damage?	?	?
Scanning acoustic microscopy	one paper	?	minutes?? but low throughput			might be about as precise as staining?	might be confident?	?	mild fixation, maybe not needed	mild fixation, maybe not needed
Acoustic sorting	not applied to chrs	??	minutes or seconds??	hasn't been shown		?	might be confident?	shear?		
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins

Indirect methods										
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Diploid culture		buffer for ~1 week	days or more	(does tell euploidy)		diploid		mostly fine but accumulates mutation?	disrupts	
Haploid culture	a few papers; logically should work	? needs eggs; low-quality ok	days or more	(does tell euploidy)		haploid; combines well with number ID		mostly fine but accumulates a bit of mutation?	disrupts; unsure how much	
Sequencing post-meiot RNA	1 paper; logically should work	? hard to develop	a day or so?	(does tell euploidy)		haploid; combines well with number ID				
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Chromosor complemer identificati	not shown; logically should work	? hard to develop	minutes?		combines with set-wise homolog ID	1				
Meiotic complemer identificati	1 paper; logically should work	medium unit price??	days?	(does tell euploidy)		haploid; combines well with number ID	much less confident with < 3 meiotic cousins		(creates it; or else disrupts)	(creates them)
Targeted complemer elimination	?	?	hours?		with more design work?	1		probably ok?	? may require fixation like for CRISPR	? may require fixation like for CRISPR
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins

Table of best methods by purpose

This table summarizes my guesses about which identification methods are most likely to work reasonably well for each functional role.

Function	Constraint	Requirements	Best guesses at workable methods		
chromosome number identification	destructive	must be high-confidence; must be cheap (for complementary identification)	PCR and low-pass sequencing (or SNP arrays)	24-FISH (or CRISPR-dCas9)	G-banding
	non-destructive	must be non-destructive; must be high-confidence	complementary identification	DAPI staining without digestion	Raman spectroscopy
	non-destructive, partial	must be non-destructive; should be cheap; can leave uncertainty (for prescreening test)	visual	electrokinesis	

Function	Constraint	Requirements	Best guesses at workable methods	
chromosome homolog identification	destructive, individual	can be medium-confidence; more information helps quantitatively	PCR and low-pass sequencing (or SNP arrays)	MFISH (or CRISPR-dCas9)
	non-destructive, setwise	must be non-destructive; can be medium-confidence	oocyte-like haploid culture	cytoplasm RNAseq (haploid)
	non-destructive, individual	must be non-destructive and high-confidence	combine non-destructive setwise homolog identification with non-destructive individual number identification	

Direct chromosome identification methods

This section gives detailed information about direct methods. The first subsection summarizes takeaways, and the following sections go into each method or type of method.

The methodology I used is basically to look around for a while, trying to find papers that apply a given direct identification method (ideally to untreated human metaphase chromosomes, or to any chromosomes, or any similarly-sized particles).

Summary of direct identification methods

There are many plausible ways to identify chromosomes by directly interacting with them, and many such methods have been investigated in the literature.

Some of these methods are very standard, e.g. staining, sequencing, and FACS. But for purposes of chromosome selection, we have different criteria than usual karyotyping. Specifically, we require a non-destructive method (or a direct method that can be used within an indirect identification method). On the other hand, we might have a less stringent requirement in that we could be satisfied with confident but very coarse genotype information—e.g. *just* chromosome number calls, or coarse crossover identification. In contrast, usually scientists would be interested in maximizing information (full DNA sequence, hetero-/eu-chromatin structure, chromosomal abnormalities, physical location, epigenetic state, etc.), and therefore would overwhelmingly favor sequencing, and might opt for fluorescent marking or similar when they want to visualize live undisturbed objects in vivo.

The existing direct chromosome identification methods seem to break up into three categories:

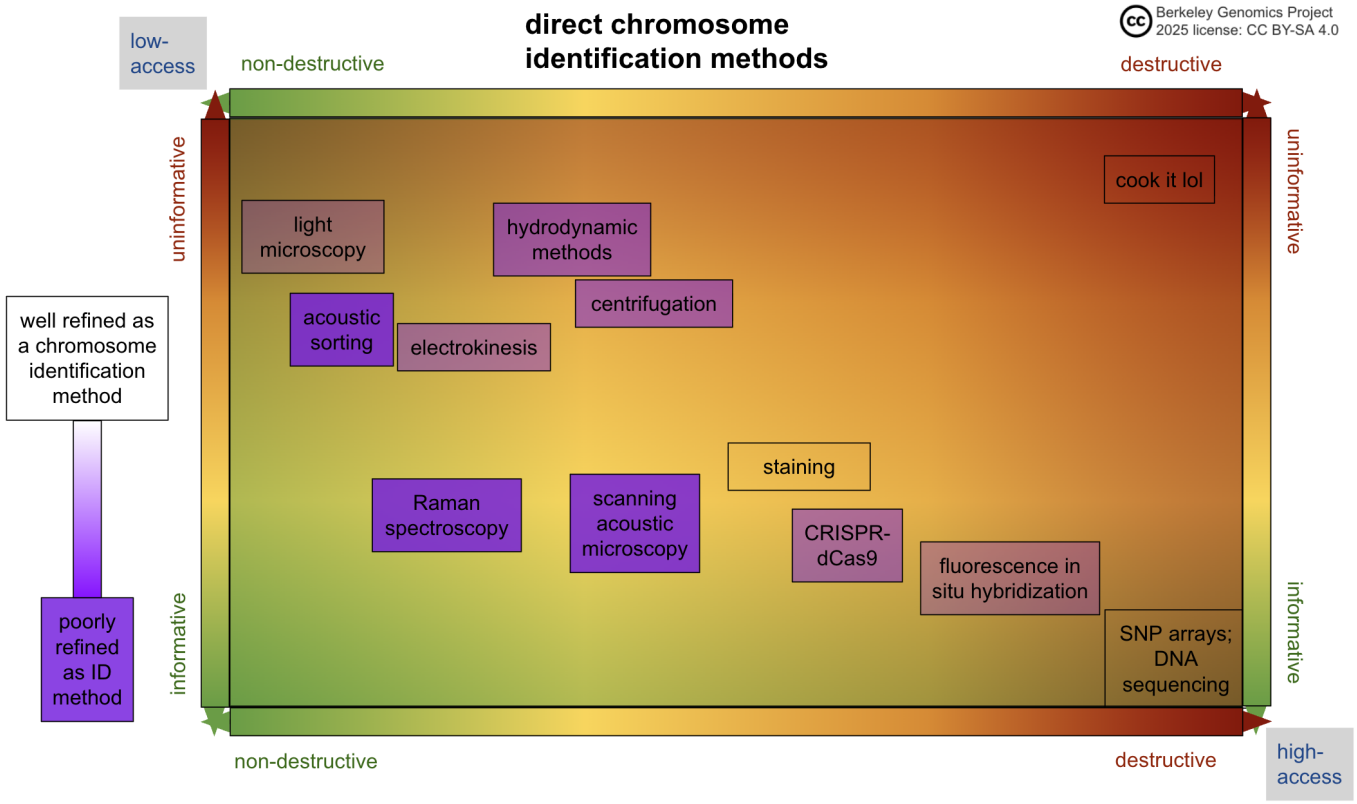
- DNA interaction.
  - Sequencing, FISH, CRISPR-dCas9, staining, etc.
  - These methods tend to be well-refined.

  - These methods tend to have a direct tradeoff: either you don't have access to the DNA, and therefore you don't get much information; or else you do have access, but the way you have access is by digesting or removing some of the chromosome's proteins, and maybe denaturing the DNA. Depending on how you remove proteins and bind to DNA, you may disrupt epigenetic state, make the chromosome vulnerable to shear damage, and even break the DNA itself.
- Gross mechanical interaction.
  - Centrifugation, hydrodynamic methods, force microscopy.
  - These methods tend to be poorly refined, as well as very imprecise and likely damaging.
- Action at a distance.
  - Acoustic methods, light microscopy, Raman spectroscopy, electrokinesis.
  - These methods are again poorly refined and quite imprecise; on the other hand, these methods are usually probably much less damaging.
  - In principle, these kinds of methods seem like they might be able to confidently, perfectly non-destructively, inexpensively, quickly identify chromosomes at least by number. But that has not been demonstrated convincingly yet, to my knowledge.

We might hope to find some workable direct method because we have different criteria than are usually pursued by scientists. However, in short, what we find is that—despite tantalizing preliminarily demonstrated possibilities—there are not currently any methods for identifying chromosomes by number and/or by homolog that meet all three of these criteria:

- the method is **well refined**, so that it is relatively inexpensive, easy, and reliable;
- the method is **non-destructive**, so that the DNA is at least intact, and ideally the epigenomic state is undisturbed;
- and the method is **highly informative / confident**, so that single chromosomes or small sets of chromosomes have their karyotype confidently called.

We can summarize these tradeoffs with a diagram (copied from the synopsis):



For this reason, we'll use indirect methods (see the section "Indirect chromosome identification methods"). It would be less complicated to use a direct identification method; indirect methods have to separately handle the two (or more) partitions of chromosomes, and apply distinct operations to them. But, it might be unavoidable to use indirect methods because of the conflicts just described.

For the purposes of indirect methods, we want a direct method that is confident, well-developed, and inexpensive. The method can be destructive, and it can even be unreliable; we can lose samples, as long as we *often* become confident that we have isolated an intact, known chromosome. (Reliability would be a major issue in strange applications such as oocyte chromosome selection, in the current regime where

oocytes are very scarce and precious. A medium case could be embryo chromosome selection, where there is a fair amount of material but it is still pretty scarce.)

Separately, we could want screening methods. (See the subsection "Screening and confirmation tests".) A direct method for screen should be fast, cheap, non-destructive, and as informative as possible—but any degree of informativeness could be useful. (The usefulness is quantitative (trying to save some amount of work on the confident-verification step) rather than binary (confidently getting 23 correctly numbered chromosomes).)

See the above section "Synopsis and takeaways" for overall context.

See the subsections "Direct and indirect chromosome identification" and "Combining multiple identification methods" for context on how direct methods might be used.

See the previous subsection "Table of best methods by purpose" for my guesses about which methods (mainly direct) are best suited for various functions.

Reprinting the summary table of judgements about each method from "Table of most methods" (see the subsection "Criteria for identification methods" for the meaning of the criteria in this table):

Direct methods										
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Light microscopy	not much applied to chrs		seconds or minutes			ambiguous sets up to 4+				
Staining		labor	hours/ 1 day			some stains roughly give singles			some fixing methods disrupt histones	fixation methods loosen proteins
Fluorescen in situ hybridizat		?	a day?		with more design work	1			denatures DNA, probably removes many histones	
CRISPR-d systems	refined, but not much with chrs	design work?	minutes or hours?	with design work	with more design work	1		probably ok?	? mildly denatures DNA, probably removes some histones	? fixation methods loosen proteins
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Low-covera DNA sequencing		~$100?	a day or two?			1				
SNP array genotyping		~$100?	a few days?			1				
Centrifuga		??	minutes or hours?			ambiguous sets up to ~4 by mass	bulk method; not confident	high shear forces?	?	?
Electrokin	not much applied to full chrs	??	hours?	hasn't been shown		multiple features might ID singles?	?	shear forces?	?	?
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Hydrodyna methods	not much applied to chrs	??	minutes??	hasn't been shown		could be ok using mass and shape?	?	shear forces?	?	?
Raman spectrosco	one paper	?	minutes?? but low throughput	claimed		might be precise? claimed 1 vs. 2	might be confident?	laser damage?	?	?
Scanning acoustic microscopy	one paper	?	minutes?? but low throughput			might be about as precise as staining?	might be confident?	?	mild fixation, maybe not needed	mild fixation, maybe not needed
Acoustic sorting	not applied to chrs	??	minutes or seconds??	hasn't been shown		?	might be confident?	shear?		

The following subsections go into more detail.

Light microscopy

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Light microscopy	not much applied to chrs		seconds or minutes			ambiguous sets up to 4+				

Is it possible to identify chromosome number just from light microscopy images?

If that were possible, it would be useful because it would be fast, non-destructive, and probably simple and inexpensive.

Banded chromosomes

There's lots of work using machine learning on images of metaphase chromosomes to automatically karyotype genomes (i.e. say what chromosomes are there, and maybe identify abnormal chromosomes). See for example the review by Sathyan et al. (2022)[14]. These systems tend to get high accuracy, calling chromosome numbers with accuracy well over 95%. But! These systems generally classify chromosomes from images of chromosomes that have been *stained* to reveal DNA bands:

[14]Remani Sathyan, Remya, Gopakumar Chandrasekhara Menon, Hariharan S, Rakhi Thampi, and Jude Hemanth Duraisamy. 'Traditional and Deep-Based Techniques for End-to-End Automated Karyotyping: A Review'. Expert Systems 39, no. 3 (2022): e12799. https://doi.org/10.1111/exsy.12799

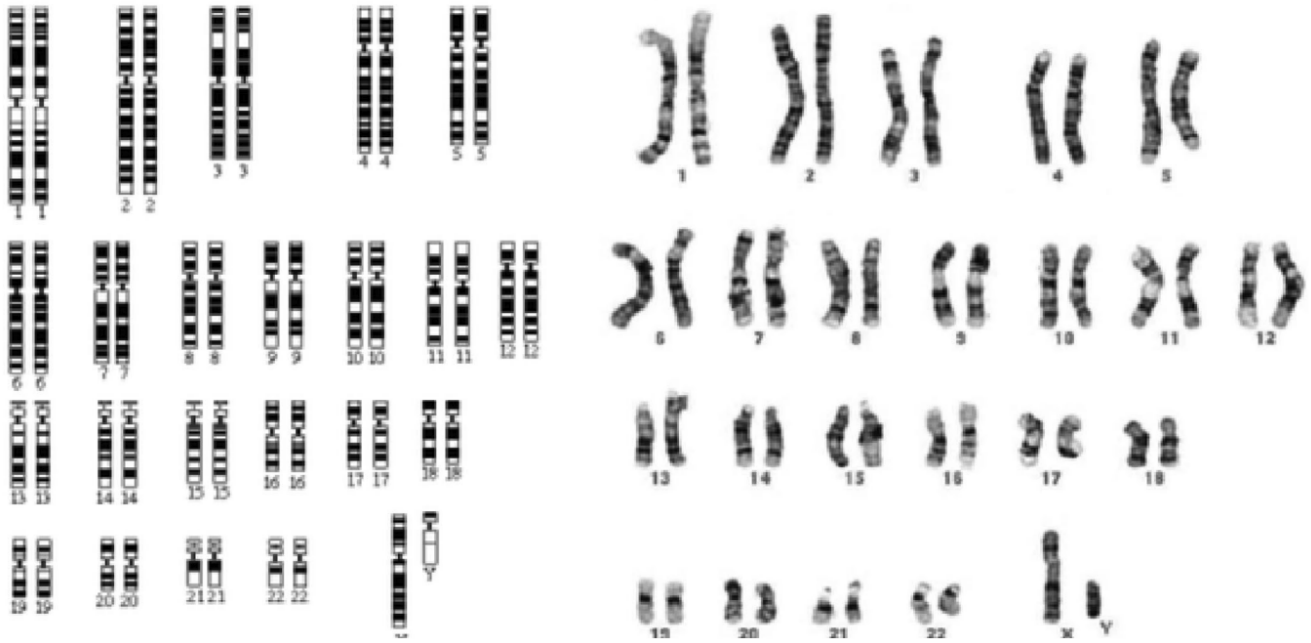


FIGURE 2 Idiogram (left) and karyogram (right)

(Figure 2 from Sathyan et al.[15]. © publisher)

These methods can be useful to speed up chromosome numbering by offloading to the computer the work of parsing banding patterns. But this is a supplement to staining methods (see the subsection "Staining"), not a standalone chromosome number identification method. Since banding patterns reveal a bunch of information characteristic of each chromosome number, the success of machine learning for these images doesn't necessarily say much about the possibility of numbering based on unstained chromosomes.

Length and centromere location

Different chromosomes have different lengths:

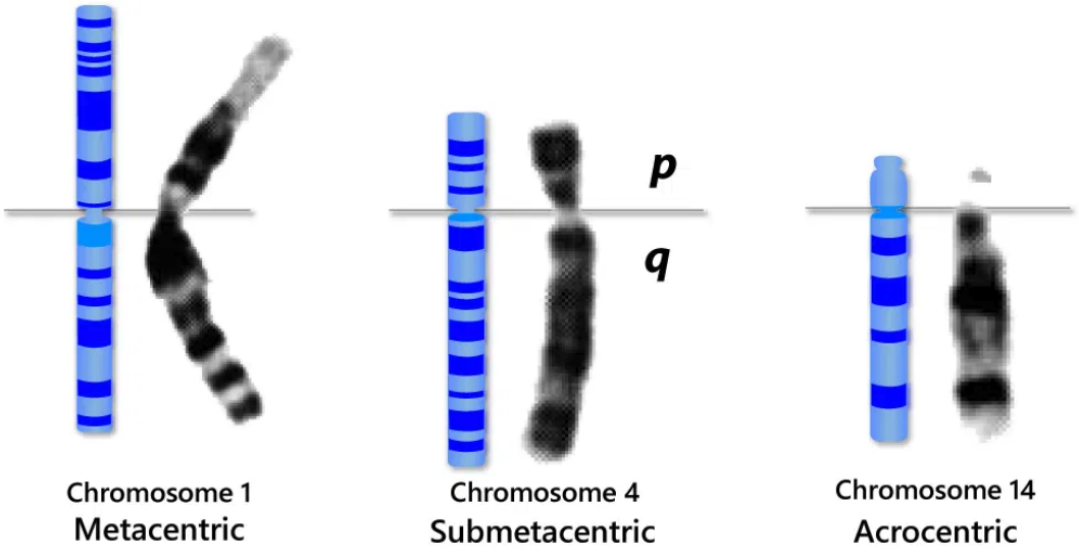
[15]Remani Sathyan, Remya, Gopakumar Chandrasekhara Menon, Hariharan S, Rakhi Thampi, and Jude Hemanth Duraisamy. 'Traditional and Deep-Based Techniques for End-to-End Automated Karyotyping: A Review'. Expert Systems 39, no. 3 (2022): e12799. https://doi.org/10.1111/exsy.12799



Figure 9.4 Giemsa-banded karyotype of a male cell.

(Figure 9.4 from Mikhail (2019)[16]. © publisher)

If you look at that image closely, you can see a major constriction spot on most of the chromosomes, where the blob becomes narrower. These are the centromeres (the part of the chromosome that gets pulled around during cell division). Different chromosome numbers have their centromeres in different positions along the length of the chromosome: metacentric (in the middle); submetacentric (off-center); or acrocentric (close to an end). (Normal human chromosomes aren't telocentric, with a centromere at the very end, but some other species have that.)



[16]Mikhail, Fady M. 'Chromosomal Basis of Inheritance∗'. In Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition), edited by Reed E. Pyeritz, Bruce R. Korf, and Wayne W. Grody. Academic Press, 2019. https://doi.org/10.1016/B978-0-12-812537-3.00009-3

(Figure from https://learn.genetics.utah.edu/content/basics/readchromosomes/[17]. © publisher)

We can classify chromosomes by approximate length and approximate centromere location. For example, here are the acrocentric chromosomes:

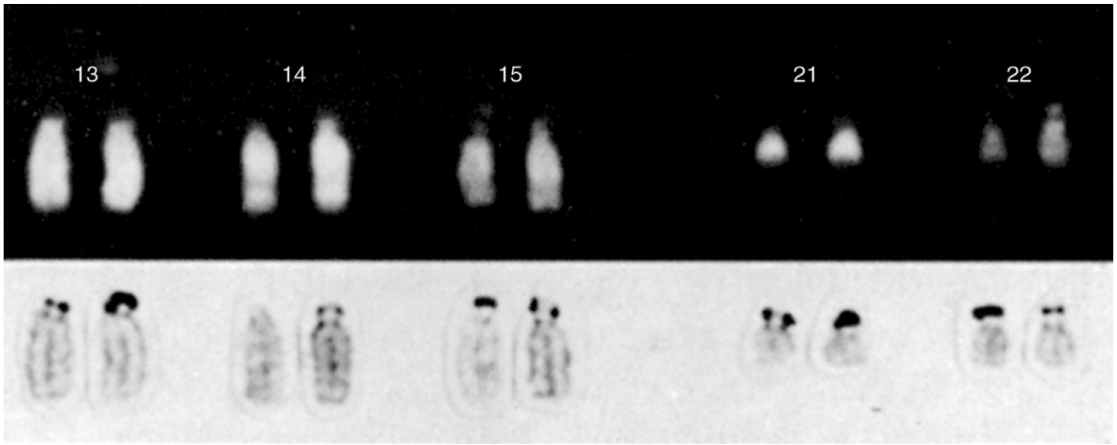


Figure 9.6 Partial karyotype of the acrocentric chromosomes from a single cell stained first by quinacrine-banding to identify each chromosome (*top*), and then by the silver NOR technique (*bottom*) to show the sites of the rRNA genes (nucleolar organizing regions, or NORs).

(Figure 9.6 from Mikhail (2019)[18]. © publisher)

In fact, it's possible to automatically coarsely classify chromosomes using just length and centromere position. Lerner et al. (1995) were able to classify chromosomes into five sets (I'm not sure of the meaning of the five sets) from images of stained chromosomes[19]. Notably, using only the inferred length and centromere position, they apparently could coarsely classify chromosomes this way with accuracy greater than 90% using just these two features:

[17]'How Do Scientists Read Chromosomes?' Accessed 15 September 2025. https://learn.genetics.utah.edu/content/basics/readchromosomes/

[18]Mikhail, Fady M. 'Chromosomal Basis of Inheritance∗'. In Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition), edited by Reed E. Pyeritz, Bruce R. Korf, and Wayne W. Grody. Academic Press, 2019. https://doi.org/10.1016/B978-0-12-812537-3.00009-3

[19]Lerner, B., H. Guterman, I. Dinstein, and Y. Romem. 'Medial Axis Transform-Based Features and a Neural Network for Human Chromosome Classification'. Pattern Recognition 28, no. 11 (1995): 1673–83. https://doi.org/10.1016/0031-3203(95)00042-X

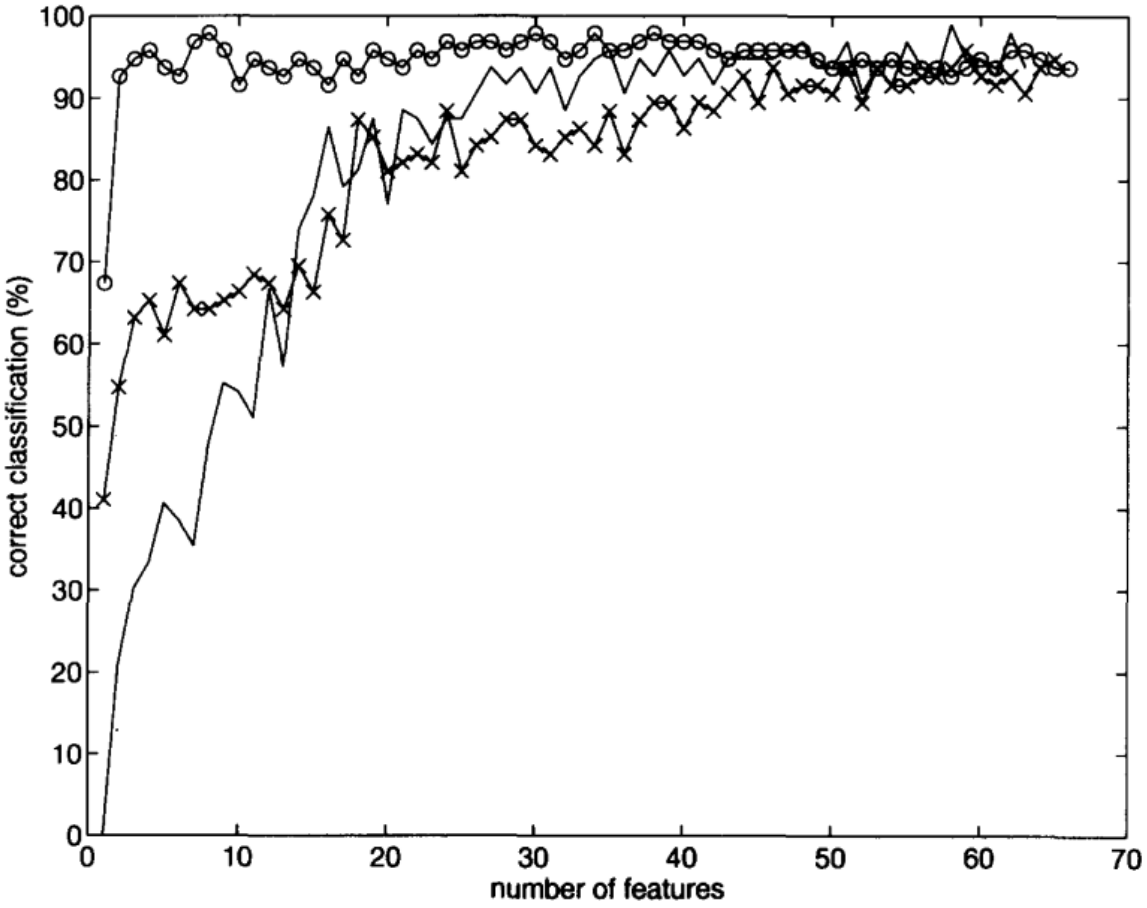


Fig. 10. The probability of correct test set classification vs the number of the best features using three feature sets: d.p. (solid line), d.p. + c.i. ("x") and d.p. + c.i. + lng ("o").

(Figure 10 from Lerner et al.[20]. Note the datapoint in the upper left, indicating >90% accuracy using two features, which are length and centromeric index. © publisher)

As another example, Ojeda et al. (2006) state "After allowing the channels to clear, chromosomes 1, 2, or 3 were screened based on size and centromere location.", where these chromosomes have not been stained. In the upper left of this image, we can see an image that might be the sort of image they were seeing when they sorted the chromosomes by size and centromere:

[20]Lerner, B., H. Guterman, I. Dinstein, and Y. Romem. 'Medial Axis Transform-Based Features and a Neural Network for Human Chromosome Classification'. Pattern Recognition 28, no. 11 (1995): 1673–83. https://doi.org/10.1016/0031-3203(95)00042-X

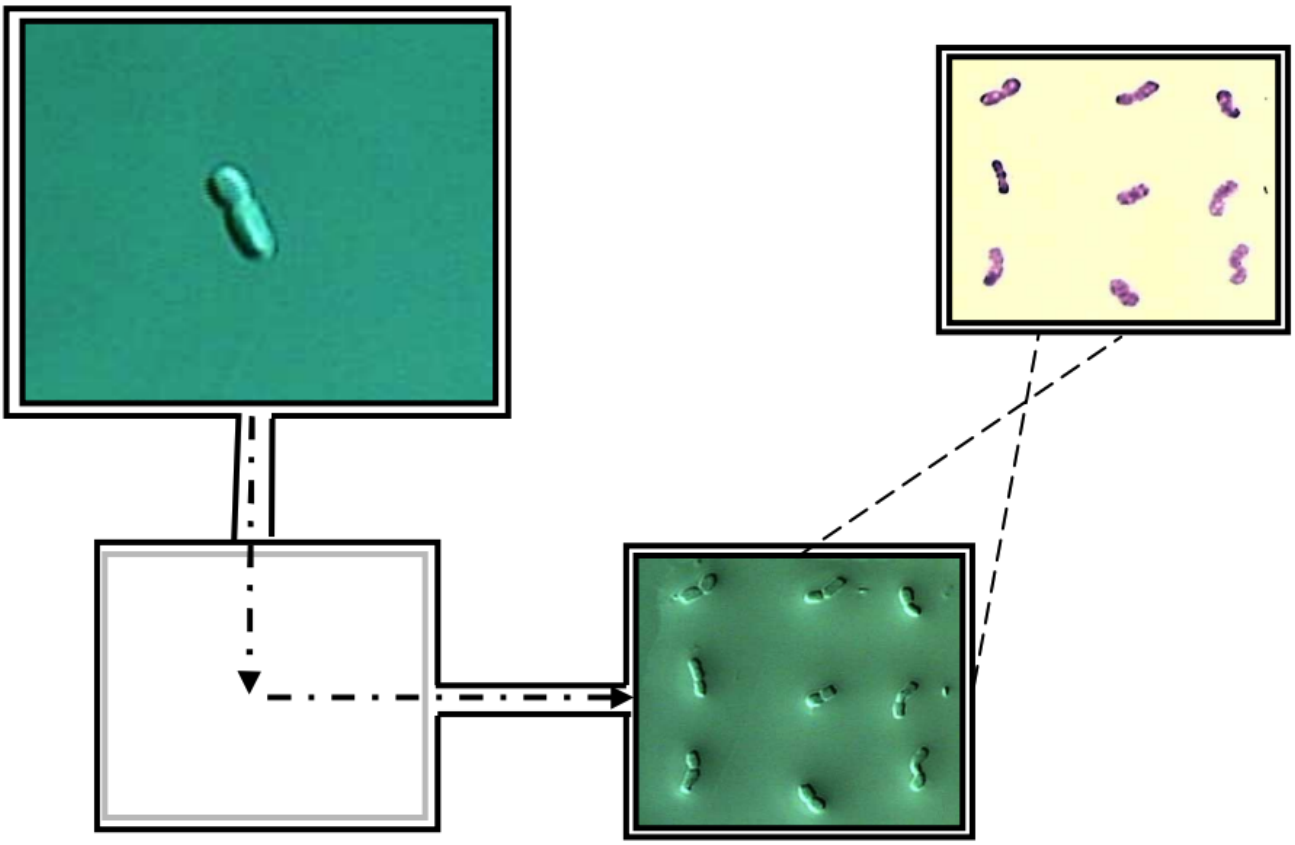


Fig. 1. Illustration of the Raman sorter for this study. The slide has three wells, one large and two smaller. The chromosomes start out in the large well and are transported using the optical tweezers to the third well, where they are fixed to the slide and then G-banded. A cytogeneticist then identified the chromosome number when possible. The chromosomes here were identified as follows: from left to right: Top row - #2, nv, #2; Middle row – nv, #2, #3; Bottom row – nv, #2, #3 (nv - not verifiable).

(Figure 1 from Ojeda et al.[21]. © publisher)

Speculations about visual identification

- It seems very unlikely that we could confidently call all exact chromosome numbers just by visual sensing. But, it's not necessary to do so; visual sorting can be useful in combination with other chromosome identification methods. (See the subsection "Combining multiple identification methods".)
- In addition to noise from measurement (low resolution imaging, obstructions, different 3D orientation, etc.), there's also actual variability in the physical shape of chromosomes. Chromosomes of the same number that come from different people, from different cell types, from different stages of cell growth, and potentially even from different cells of the same type, might have different lengths and thicknesses. These variations certainly don't make all chromosomes visually ambiguous with each other; but they may put significant practical upper bounds on how well you can discern some sets of chromosome numbers with similar lengths[22]. I don't know whether there are unresolvable ambiguities; the situation isn't simple because there are multiple large-scale features (length, thickness, centromere position) that could hypothetically be used for visual identification.
- It might be possible to leverage modern big image data and pretraining to help address the issue of scarce data. For example, maybe we could take a large pretrained image model and obtain from it some fairly low-dimensional embedding space. Then we could train a chromosome number classifier on pairs of the form (latent vector for visual image; ground truth from staining).

[21]Ojeda, Jenifer F., Changan Xie, Yong-Qing Li, Fred E. Bertrand, John Wiley, and Thomas J. McConnell. 'Chromosomal Analysis and Identification Based on Optical Tweezers and Raman Spectroscopy'. Optics Express 14, no. 12 (2006): 5385–93. https://doi.org/10.1364/OE.14.005385

[22]Van Dyke, D. L., Maria J. Worsham, L. J. Fisher, and L. Weiss. 'The Centromere Index and Relative Length of Human High-Resolution G-Banded Chromosomes'. Human Genetics 73, no. 2 (1986): 130–32. https://doi.org/10.1007/BF00291602

- I don't know what sort of light microscopy is suitable for visual chromosome sorting.
- Staining *without* treating with a protease could be useful. This tends to produce solid stains, which would make imaging easier (though there would not be bands). If it is non-destructive, then it could make it easier to visualize chromosome size and shape.
- It's possible that much gentler treatments are possible. For example, since microscope cameras have been developed to be much more sensitive, we can maybe get away with using much lower concentrations of dyes. Fuchs et al. (2023) report that using Hoechst 33342 dye (a standard stain that's relatively less cytotoxic) at concentrations a couple orders of magnitude lower than standard practice was far less cytotoxic, but still feasible to image[23].
- A 1981 paper used microscope photometry to image chromosomes[24]. The chromosomes were trypsinized. But, the image at least suggests that if untrypsinized chromosomes have patterns of differing density, and those patterns characterize different chromosomes numbers, then number might be identifiable this way:

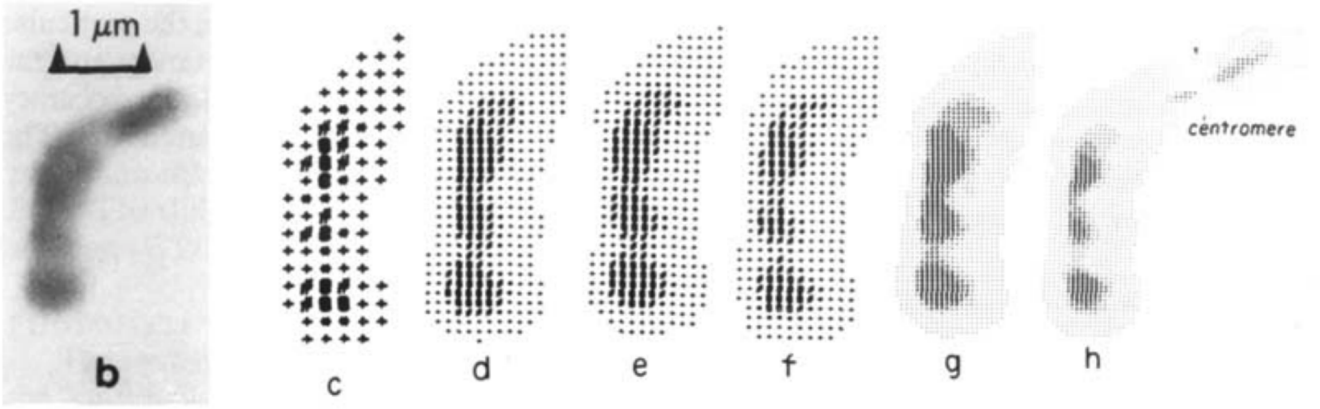


Fig. 2. Comparison of a photomicrograph of chromosome No. 15 together with photometric scans of the same member performed with varying step sizes and with different diameters of circular measuring diaphragms; percentage of area overlap between consecutive readings shown in parentheses. (b) Photomicrograph of No. 15 chromosome. Photometric scans of the same chromosome: (c) 0·2 μm stepsize and 0·2 μm diaphragm (0%); (d) 0·1 μm stepsize and 0·3 μm diaphragm (58%); (e) 0·1 μm stepsize and 0·2 μm diaphragm (39%); (f) 0·1 μm stepsize and 0·1 μm diaphragm (0%); (g) 0·05 μm stepsize and 0·16 μm diaphragm (61%); (h) 0·05 μm stepsize and 0·1 μm diaphragm (39%).

(Figure 2 from Wayne and Sharp (1981). © publisher)

Staining

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Staining		labor	hours/ 1 day			some stains roughly give singles			some fixing methods disrupt histones	fixation methods loosen proteins

The general idea of staining is to put a mass of staining molecules onto the chromosomes. The staining molecules bind to some parts of the chromosomes, lighting them up and making a visible pattern that's characteristic of a specific chromosome.

How staining works

In general, a staining protocol looks something like this:

[23]Fuchs, Heiko, Kirsten Jahn, Xiaonan Hu, Roland Meister, Maximilian Binter, and Carsten Framme. 'Breaking a Dogma: High-Throughput Live-Cell Imaging in Real-Time with Hoechst 33342'. Advanced Healthcare Materials 12, no. 20 (2023): 2300230. https://doi.org/10.1002/adhm.202300230

[24]Wayne, A. W., and J. C. Sharp. 'The Use of High Resolution Microscope Photometry in the Discrimination of Chromosome Bands'. Journal of Microscopy 124, no. 2 (1981): 163–67. https://doi.org/10.1111/j.1365-2818.1981.tb00309.x.

1. Get some M-phase chromosomes.
2. Briefly (seconds or minutes) treat the chromosomes with trypsin to remove some proteins bound to the DNA.
3. Apply the staining molecule(s).
4. Look at the stained chromosomes in a microscope, and use the pattern of light and dark patches to figure out which chromosome number you're looking at.

Protocols vary in several ways, e.g.:

- Which staining molecule is used.
- How the chromosomes are treated, e.g. with trypsin, drying, heating, etc.
- How long the protocol takes (hours to tens of hours).
- What pattern of light and dark patches is produced.
- What sort of microscope is used (light microscopy or fluorescence microscopy).

See Estandarte (2012) for a nice review of staining methods[25].

How do staining molecules bind to DNA? There are several ways:

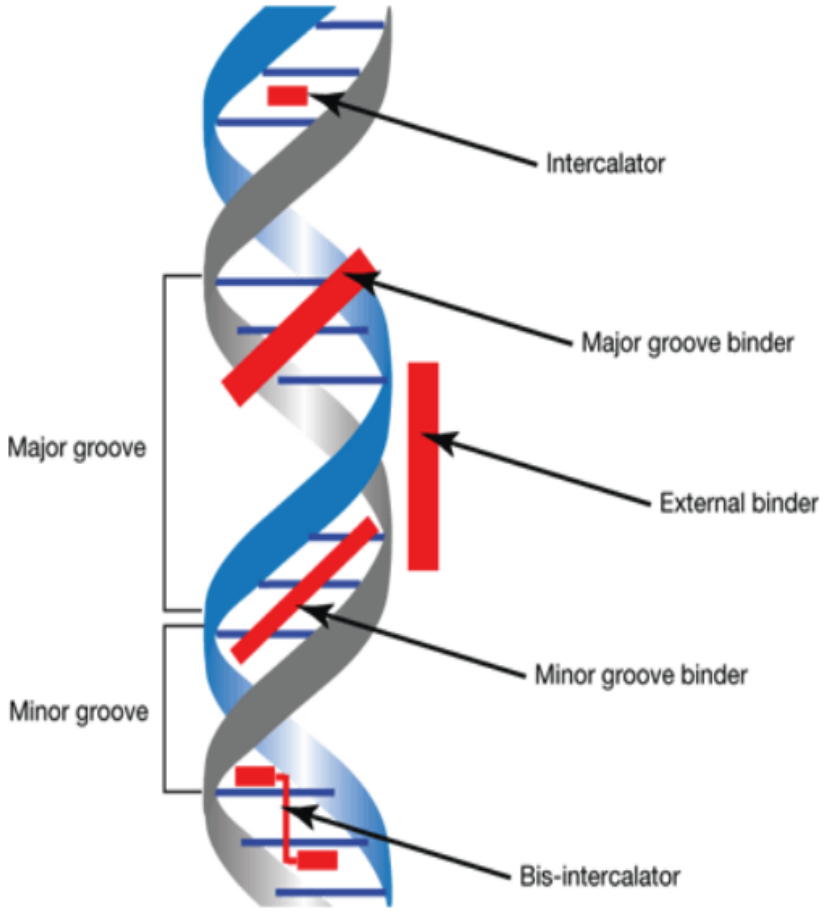


Figure 3.4. Modes of binding of a stain to DNA.

(Figure 3.4 from Estandarte (2012). © publisher)

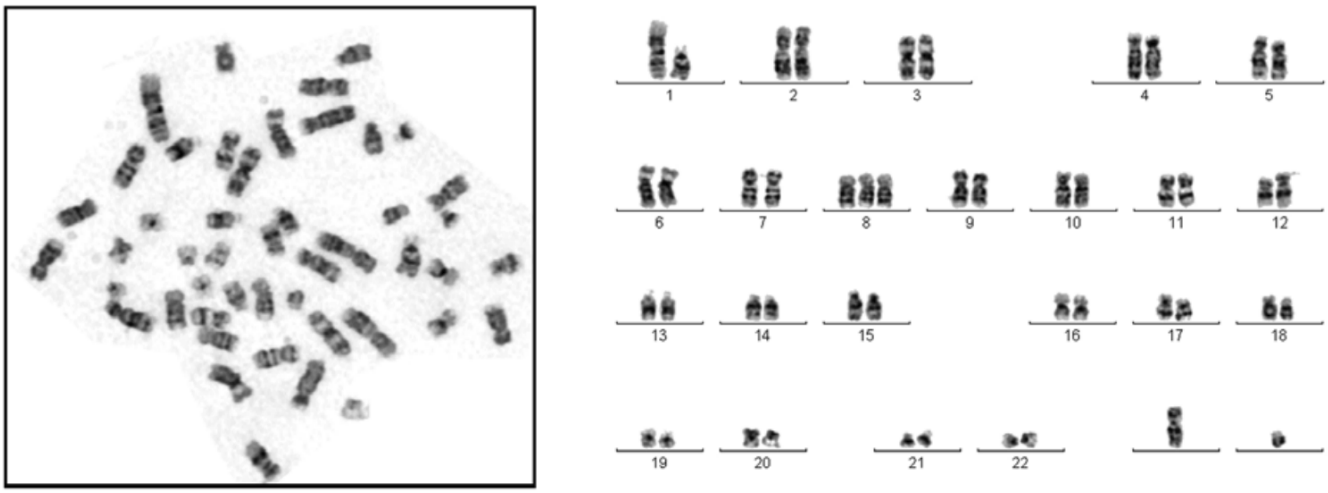
Stains with different kinds of binding have different properties. E.g. intercalating binders tend to be less sequence-specific, whereas a minor groove binder might for example preferentially bind to AT-rich regions or GC-rich regions. Intercalating binders might also disrupt DNA structure more.

What most of these staining methods have in common is that they work by denaturing the chromatin somewhat. We treat the chromatin with trypsin or another protease, and that removes some of the proteins

[25]Estandarte, A. 'A Review of the Different Staining Techniques for Human Metaphase Chromosomes'. 2012. https://www.ucl.ac.uk/~ucapikr/projects/Ana_staining_LitRev.pdf

bound to the DNA. With some of those proteins removed, the DNA is more open for the staining molecules to bind.

Different sections of a chromosome will lose more or less proteins during trypsinization. Then, the stain will bind more or less depending on how much of those proteins were extracted. That makes the banding pattern. For example, here is an image of G-banded human chromosomes, produced by protease treatment and Giemsa staining:



(Figure 22.3.1 from Bayani and Squire (2004)[26]. © publisher)

Staining for chromosome identification

Staining is fairly attractive as a chromosome number identification method:

- It has been a standard and widely practiced method. So there are standard methods that work well for various sorts of chromosomes.
- Banding provides near-certain number calls, if the treatment and imaging is high-quality. Banding patterns are generally very similar between different people and different cells.
- Although manually analyzing the images is laborious, as mentioned above in "Light microscopy", there is a lot of work on using machine learning to automatically number chromosomes from staining images, and those automatic methods seem plausibly sufficient.
- Finally, staining is pretty fast; somewhere between an hour and a day, with probably less than an hour of bench work. I'm not sure how fast it could be made if corners are cut. Scientists doing banding are often interested in more precise information than just the karyotype—e.g. they might want to know about large translocations or investigate intra-chromosome structure. So, they want clear bands. We might be able to get away with faster, lower-quality treatments, resulting in banding that's lower-quality (e.g. fuzzier) but good enough for chromosome numbering.

For these reasons, staining is a plausible direct identification method. However, there are several issues with staining:

- Most ways of staining for producing bands are probably destructive.
  - First, when you treat the DNA with a protease, proteins are removed, and this seems likely to disrupt the epigenomic state of the DNA. For example, I don't know if histones in particular tend to be removed by e.g. trypsinization, but if they are, that would disrupt histone modifications.
  - Further, stains are likely to cause DNA damage. I don't know this with certainty, but for example, even a Hoechst stain for live cells tends to be toxic to the cells. In particular, Zhao et al. (2009) found that Hoechst 33342, a standard fluorescent dye, as well as some other dyes, induce a response from cells that looks like the cell's response to DNA damage[27]. It might be possible to get an adequate stain for chromosome numbering with much smaller amounts of dye[28], which

[26]Bayani, Jane, and Jeremy A. Squire. 'Traditional Banding of Chromosomes for Cytogenetic Analysis'. Current Protocols in Cell Biology 23, no. 1 (2004): 22.3.1-22.3.7. https://doi.org/10.1002/0471143030.cb2203s23

[27]Zhao, Hong, Frank Traganos, Jurek Dobrucki, Donald Wlodkowic, and Zbigniew Darzynkiewicz. 'Induction of DNA Damage Response by the Supravital Probes of Nucleic Acids'. Cytometry. Part A : The Journal of the International Society for Analytical Cytology 75, no. 6 (2009): 510–19. https://doi.org/10.1002/cyto.a.20727

[28]Fuchs, Heiko, Kirsten Jahn, Xiaonan Hu, Roland Meister, Maximilian Binter, and Carsten Framme. 'Breaking a Dogma: High-Throughput Live-Cell Imaging in Real-Time with Hoechst 33342'. Advanced Healthcare Materials 12, no. 20 (2023): 2300230. https://doi.org/10.1002/adhm.202300230

might avoid DNA damage, but that is just speculation. Sen et al. (2018) found that even a reportedly mild dye (SiR-Hoechst) interfered with cell processes[29]. It could be that a given stain doesn't inherently directly damage DNA, but it causes DNA damage by interfering with some cell processes, or it is otherwise cytotoxic. Chromosomes labeled with that stain cannot safely be part of a genome to be used for zygogenesis, because the resulting zygote would be prone to having a damaged genome.

- Another limitation is that in most cases (barring large-scale chromosome abnormalities), staining methods can't distinguish homologs.
- Finally, it might be difficult and/or unreliable to get a good stain. If that's right, then for purposes of complementary identification, it's effectively lower-confidence.

So, in general, staining is a direct, destructive, fast, maybe-confident chromosome number identification method.

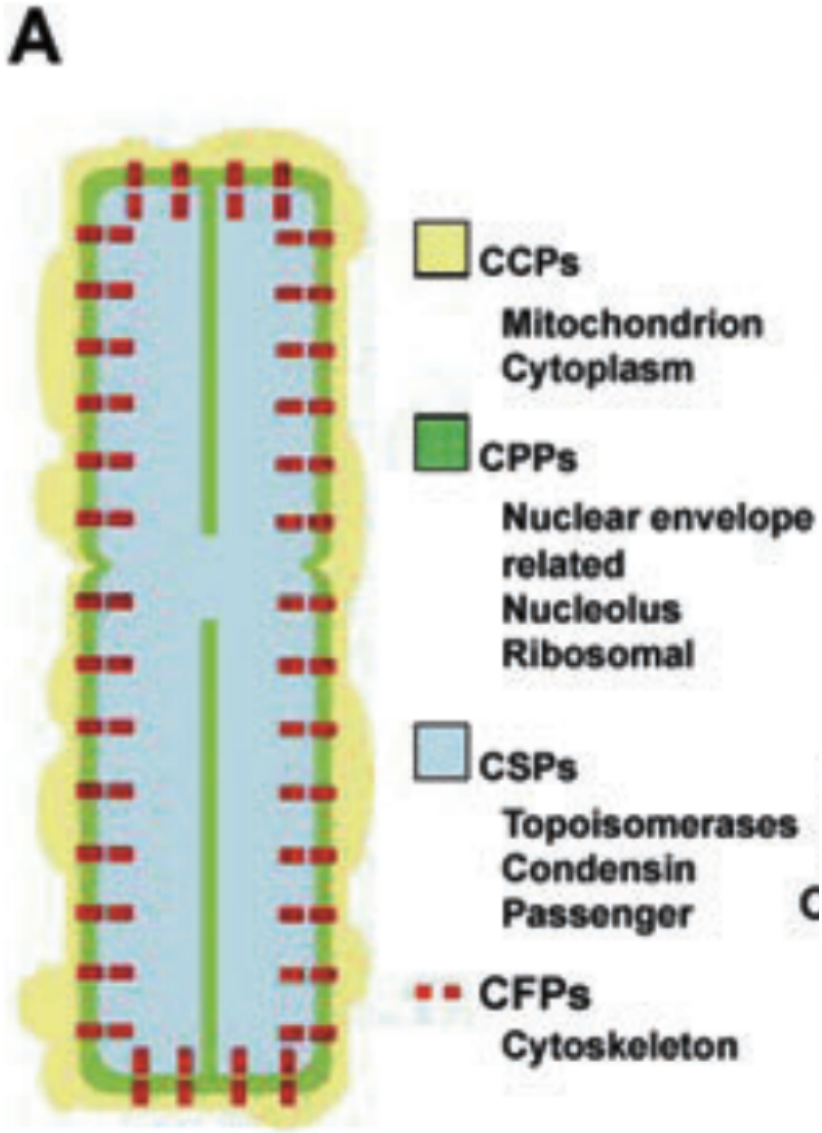
Immunostaining chromosome structural proteins

A metaphase chromosome contains not just DNA, but also hundreds of different kinds of proteins. Histones are key proteins that form nucleosomes, which the DNA wraps around. There are also many other proteins, many of which provide structure to the chromosome.

Fukui and Uchiyama (2007) analyzed the protein content of human metaphase chromosomes[30]. "Coating" proteins are on the outside, but maybe aren't very strongly bound to the surface and aren't essential. The next layer is made of "peripheral" proteins, which are more structural. This suggests a coarse layered model of a metaphase chromosome with its proteins:

[29]Sen, Onur, Adrian T. Saurin, and Jonathan M. G. Higgins. 'The Live Cell DNA Stain SiR-Hoechst Induces DNA Damage Responses and Impairs Cell Cycle Progression'. Scientific Reports 8, no. 1 (2018): 7898. https://doi.org/10.1038/s41598-018-26307-6

[30]Fukui, Kiichi, and Susumu Uchiyama. 'Chromosome Protein Framework from Proteome Analysis of Isolated Human Metaphase Chromosomes'. The Chemical Record 7, no. 4 (2007): 230–37. https://doi.org/10.1002/tcr.20120.

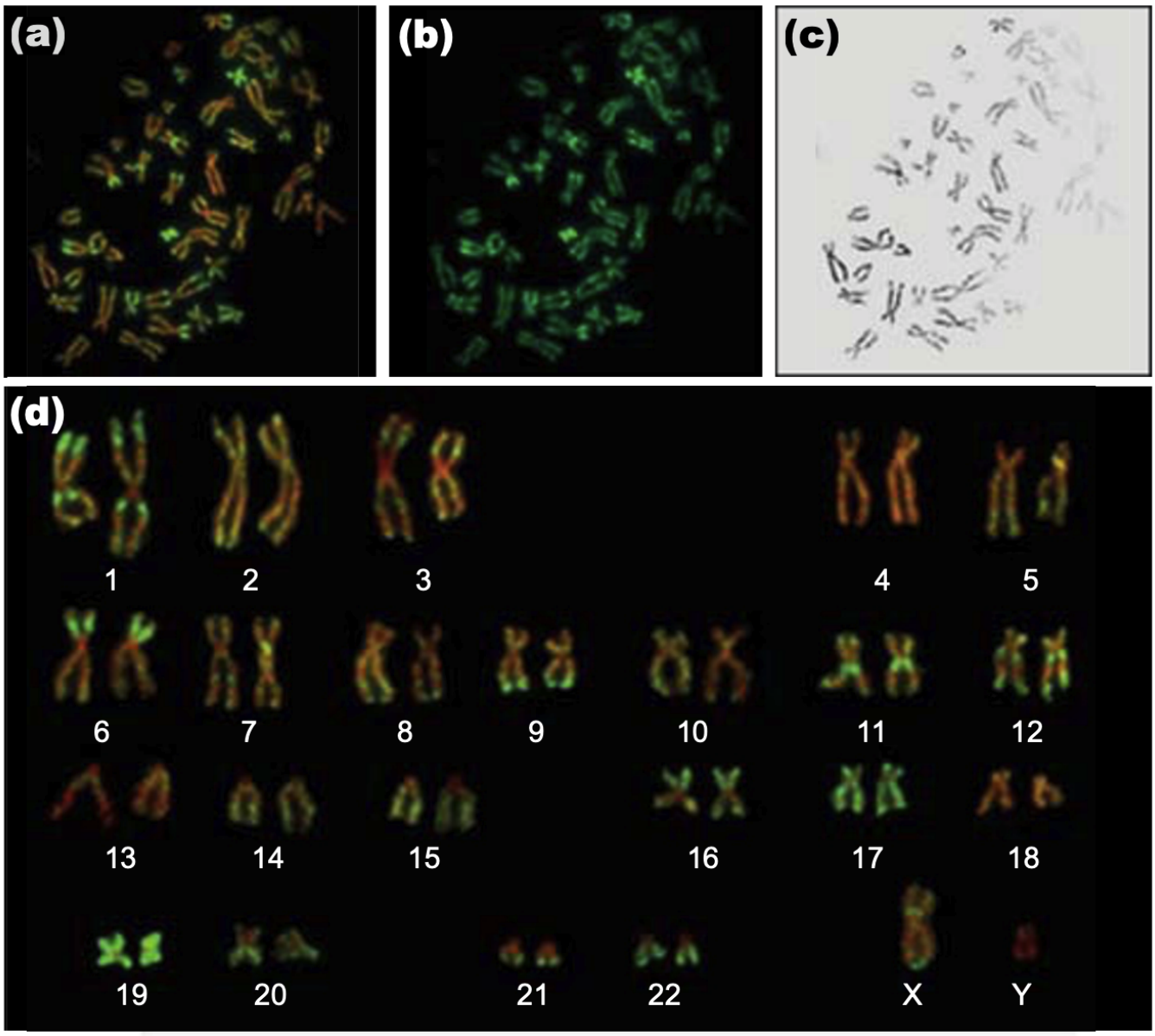


(Figure 5 from Fukui and Uchiyama (2007). © publisher)

In theory, there might be a stain that would bind to chromosome peripheral proteins of truly untreated metaphase chromosomes; and that might reveal some kind of banding pattern. I don't know what's been tried here, but standard stains presumably wouldn't work because if they did then people wouldn't do deproteinating when trying to get a karyotype. My guess is that banding patterns are always created by patterns of macro-scale G/C vs. A/T density in the DNA, either directly via stains that bind to those pairs or indirectly via stains that bind to histones (which are denser or sparser depending on GC content). In that case, you don't get banding patterns by staining peripheral proteins; unless the physical volume is affected by histone density, in which case variations in physical volume (say, chromosome diameter) might be detectable this way.

I don't know if DNA or histones are exposed in untreated chromosomes. Terrenoire et al. (2010) stained unfixed chromosomes with DAPI and with histone antibodies[31]. They state that they were then able to karyotype the resulting set of chromosomes:

[31]Terrenoire, Edith, Fiona McRonald, John A. Halsall, et al. 'Immunostaining of Modified Histones Defines High-Level Features of the Human Metaphase Epigenome'. Genome Biology 11, no. 11 (2010): R110. https://doi.org/10.1186/gb-2010-11-11-r110.



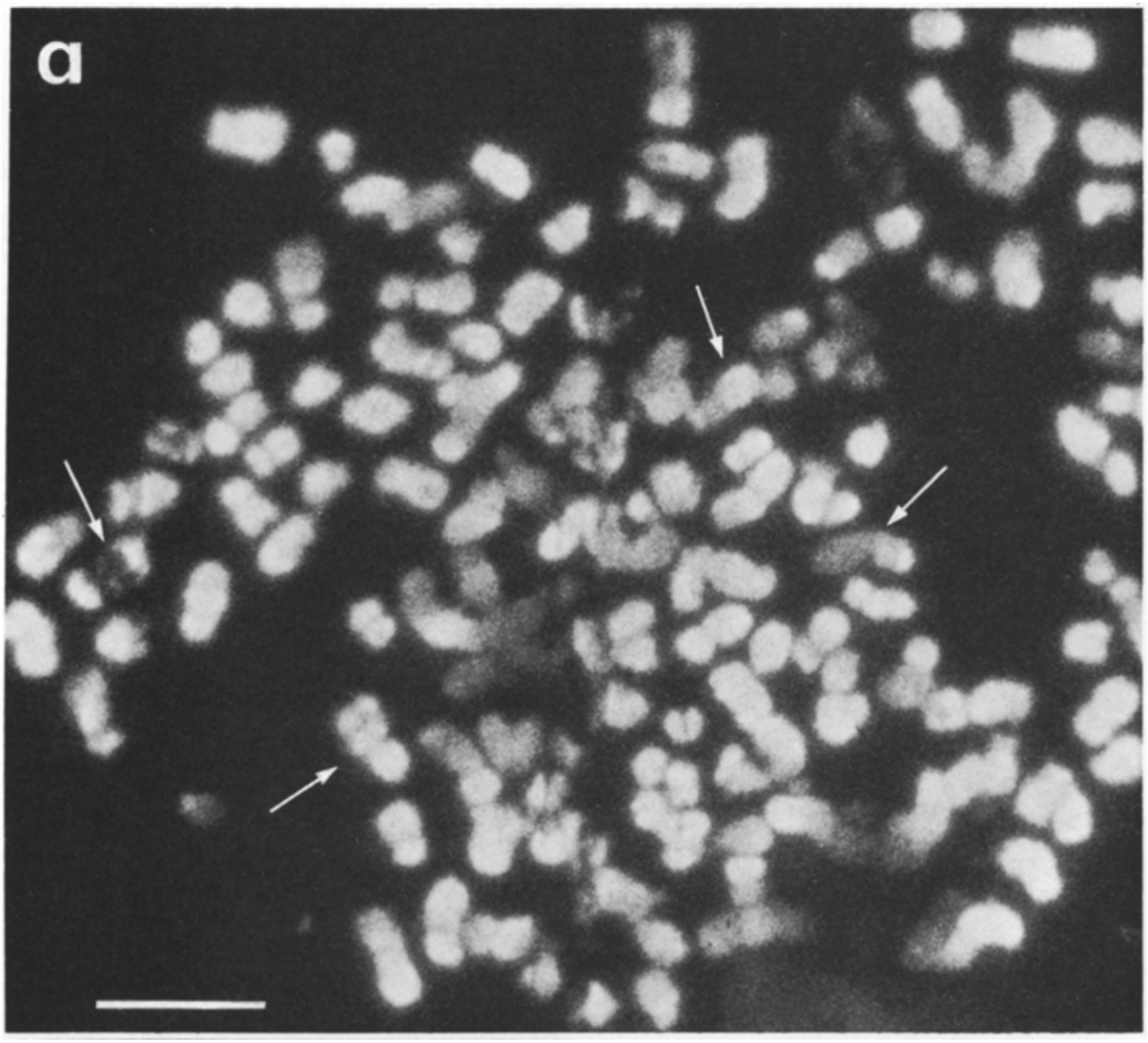
(Figure 1 from Terrenoire et al. (2010). Red is DAPI, green is a H3K4me3 antibody. © publisher)

However, their protocol, adapted from Jeppesen et al. (1992)[32] and others, does treat the chromosomes with Triton X-100 detergent, which removes some proteins, though probably less than trypsin or other proteinases.

Turner (1982) also used histone antibodies to stain human metaphase chromosomes[33]. That stain got some hints of banding (see arrows):

[32]Jeppesen, Peter, Arthur Mitchell, Bryan Turner, and Paul Perry. 'Antibodies to Defined Histone Epitopes Reveal Variations in Chromatin Conformation and Underacetylation of Centric Heterochromatin in Human Metaphase Chromosomes'. Chromosoma 101, no. 5 (1992): 322–32. https://doi.org/10.1007/BF00346011.

[33]Turner, Bryan M. 'Immunofluorescent Staining of Human Metaphase Chromosomes with Monoclonal Antibody to Histone H2B'. Chromosoma 87, no. 3 (1982): 345–57. https://doi.org/10.1007/BF00327635.



(Figure 2 from Turner (1982). © publisher)

Note that the discolorations are the same on each of the two paired sister chromatids, so it's not just random. Anyway, these chromosomes were even more untreated—just ethanol. Still, ethanol also denatures and loosens proteins, so it's still not completely clear whether or not the histones were exposed with truly untreated chromosomes. (Indeed, Turner notes that some chromosomes only stained faintly, but that this was corrected by soaking the chromosomes in salt to further loosen the protein coating.) Logically, one might be able to loosen chromosome proteins to expose histones to get a good-enough stain, without degrading the structure too much and making the chromosome vulnerable to shear; but I don't know if that's actually doable.

MicroSort

MicroSort is a company that offers sperm sorting by sex. They stain sperm with Hoechst 33342 and then sort by activated brightness; since an X chromosome is bigger than a Y chromosome, female sperm are slightly (~3%) brighter than male sperm. In 2009, and again in 2014, they claimed (very roughly) 90% effectiveness, i.e. roughly 90% of the sperm have the sex that they think it does after sorting[34][35]. This is fairly poor confidence, but 3% is a close margin (and methods may have improved since then). They also claim to have healthy pregnancy rates broadly comparable to rates for normal IVF[36]. But, I don't know how

[34]Karabinus, D. S. "Flow Cytometric Sorting of Human Sperm: MicroSort® Clinical Trial Update." Theriogenology 71, no. 1 (2009): 74–79. https://doi.org/10.1016/j.theriogenology.2008.09.013.

[35]Karabinus, David S., Donald P. Marazzo, Harvey J. Stern, et al. "The Effectiveness of Flow Cytometric Sorting of Human Sperm (MicroSort®) for Influencing a Child's Sex." Reproductive Biology and Endocrinology 12, no. 1 (2014): 106. https://doi.org/10.1186/1477-7827-12-106.

[36]Marazzo, Donald P., David Karabinus, Lawrence A. Johnson, and Joseph D. Schulman. "MicroSort® Sperm Sorting Causes No Increase in Major Malformation Rate." Reproduction Fertility and Development 28, no. 10 (2015): 1580–87.

safe it is; other studies do find negative effects at least on the physiology of sperm from higher concentrations of Hoechst 33342[37].

The fact that MicroSort works suggests that Hoechst is able to penetrate inside the sperm chromatin mass. If that's right, then other molecules might also be able to penetrate. On the other hand, it could be that the sorting works just by the difference in dye binding to slightly different-sized surfaces of nuclei.

If in fact Hoechst 33342 staining is safe for sperm, that suggests that it would also be safe for chromosome staining. It's conceivable that it's only safe for sperm but not for chromosomes, e.g. because there's some repair proteins in sperm that wouldn't be present for isolated chromosomes. There were concerns with the safety of MicroSort[38]; it wasn't given FDA approval and is not available in the US.

Hybridization methods

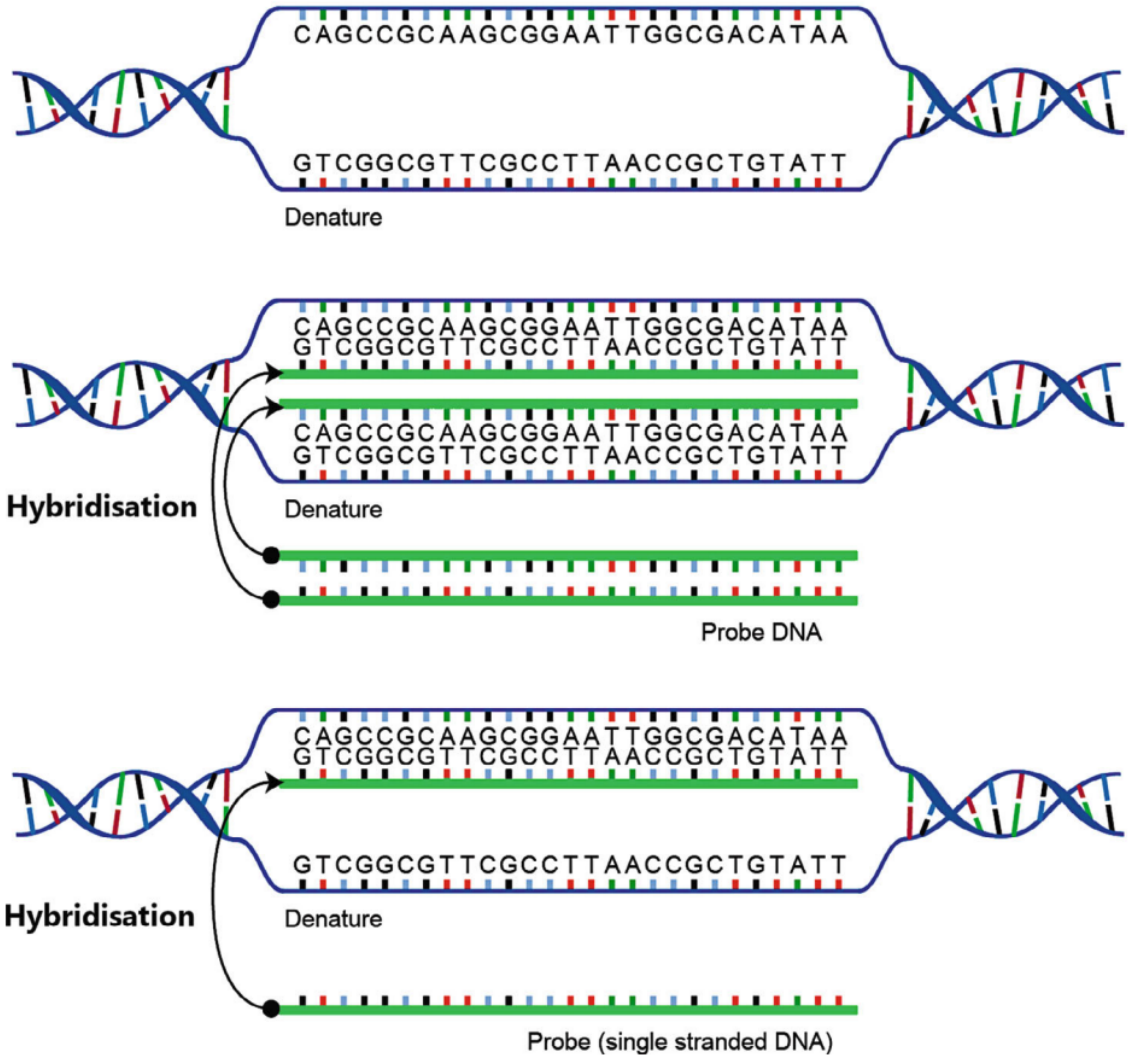
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Fluorescen in situ hybridizat		?	a day?		with more design work	1			denatures DNA, probably removes many histones	
CRISPR-d systems	refined, but not much with chrs	design work?	minutes or hours?	with design work	with more design work	1		probably ok?	? mildly denatures DNA, probably removes some histones	? fixation methods loosen proteins

Hybridization methods involve DNA hybridization. That's when double-stranded DNA (such as a chromosome) is denatured (unzipped), so that two complementary single-strand regions are exposed. Then some other single-strand DNA lines up with those regions and hybridizes (binds) to them. We somehow detect the hybridized portions of DNA.

https://doi.org/10.1071/RD15011.

[37] Quan, Guo Bo, Yuan Ma, Jian Li, et al. "Effects of Hoechst33342 Staining on the Viability and Flow Cytometric Sex-Sorting of Frozen-Thawed Ram Sperm." Cryobiology 70, no. 1 (2015): 23–31. https://doi.org/10.1016/j.cryobiol.2014.11.002.

[38] Caroppo, Ettore. 'Sperm Sorting for Selection of Healthy Sperm: Is It Safe and Useful?' Fertility and Sterility 100, no. 3 (2013): 695–96. https://doi.org/10.1016/j.fertnstert.2013.06.006.



(Figure 4.11 from Crouch (2025)[39]. © publisher)

These methods are good in that we can precisely visualize specific DNA motifs. This would enable accurate chromosome number identification, and also more precise measurements like tracking whether chromosomes that we're manipulating have gotten their ends broken off. Further, with more design work to make probes that bind to one allele out of multiple possible alleles, hybridization can detect allele differences and therefore do homolog identification. E.g. if you have 10 allele-specific probes on one chromosome, you can approximately locate crossovers to within 1/10th of the length of the chromosome.

Also, hybridization methods tend to be fairly fast; denaturation might take several hours, but the whole process can be done in under a day, which is not bad. But, even faster would be better, in order to quickly iterate on experimental setups, e.g. debugging some sorting method or training some low-confidence identification method.

One problem with hybridization methods is that they all involve denaturing the target DNA. Denaturation isn't inherently destructive; a cell functioning normally is constantly transcribing DNA by locally denaturing it, and then rezipping it back up when appropriate. But, denaturation is usually done globally, e.g. by heating up the DNA, which tends to cause damage.

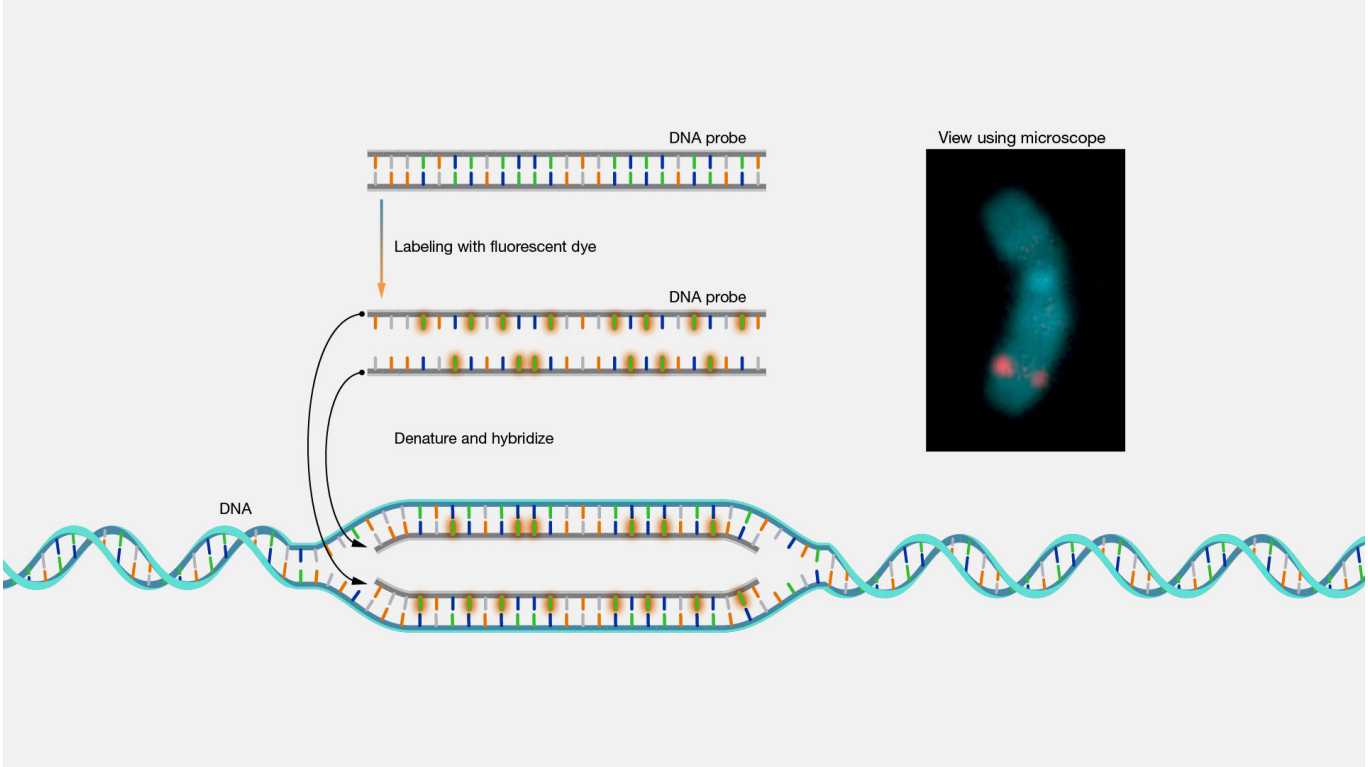
Further, hybridization requires that the target DNA be exposed, which means some proteins have to be removed. Removing proteins risks disrupting some epigenomic state.

Hybridization is useful in the context of culturing (see the subsection "Cell culturing vs. isolating-ensembling methods"). In that context, hybridization can be used to select cells based on whether or not they have the desired chromosomes (containing the targeted DNA segments), or otherwise perform operations on targeted chromosomes (see the subsection "Targeted DNA operations"). This can often be more efficient than directly sequencing and selecting cell lines.

[39]Crouch, Mair. 'Methods Used to Study the Genome'. In Medical Genetics and Law: An International Perspective, edited by Mair Crouch. Springer Nature Switzerland, 2025. https://doi.org/10.1007/978-3-031-78958-8_4

Fluorescence in situ hybridization

The idea of fluorescence in situ hybridization (FISH) is to add some DNA fragments that are the same sequence as the DNA you want to label. The DNA fragments you add are fluorescent (originally, radioactive nucleotides; now, fluorescent labels). You denature the DNA you want to label, i.e. you make it unzip into two strands (generally by cooking it). Once it's unzipped, the fluorescent fragments hybridize (bind, re-zip) together with the complementary strands of the target DNA. Since the fluorescent strands of the fragments bind specifically to their complementary DNA motif within the target DNA, you can see where the DNA is lighting up, and then you know the presence and position of those complementary motifs. An illustration:



(Figure from Dutra[40]. © publisher)

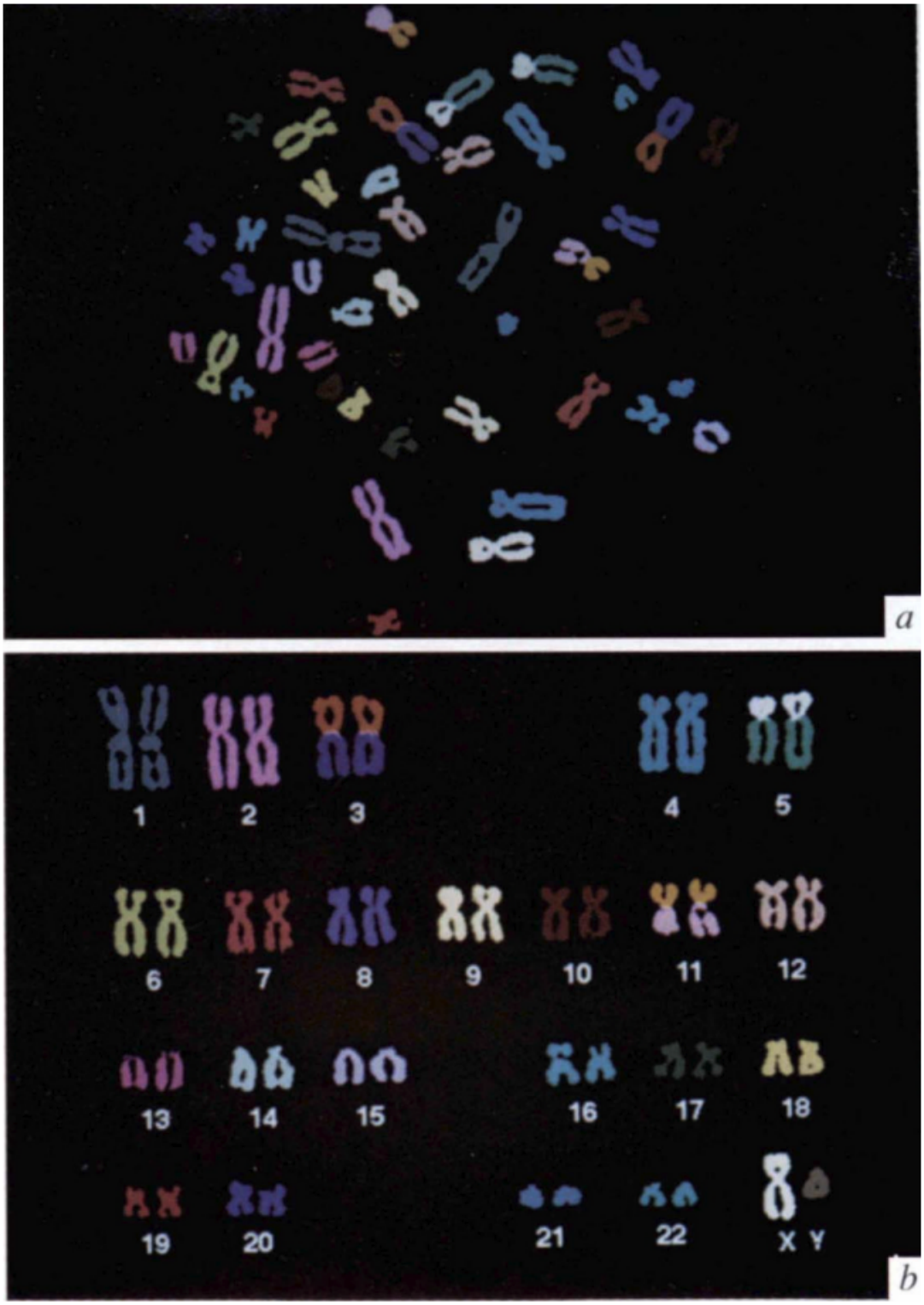
Since FISH can be targeted to specific DNA motifs, it's highly versatile. For example, it can be used to paint telomeres, which could be useful for tracking whether our chromosomes are breaking; and it can be used to paint chromosome-specific alpha satellites (repetitive DNA segments near centromeres)[41].

Multiplex-FISH uses many probes for all chromosomes, where probes for different chromosomes have different combinations of fluorophores. Then you can tell the difference between different chromosomes just by the color differences[42]. The resulting karyotype looks like this (pseudocolored):

[40]Dutra, Amalia. 'Fluorescence In Situ Hybridization (FISH)'. Accessed 16 September 2025. https://www.genome.gov/genetics-glossary/Fluorescence-In-Situ-Hybridization-FISH

[41]Dunham, Ian, Christoph Lengauer, Thomas Cremer, and Terry Featherstone. 'Rapid Generation of Chromosome-Specific Alphoid DNA Probes Using the Polymerase Chain Reaction'. Human Genetics 88, no. 4 (1992): 457–62. https://doi.org/10.1007/BF00215682

[42]Speicher, Michael R., Stephen Gwyn Ballard, and David C. Ward. 'Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH'. Nature Genetics 12, no. 4 (1996): 368–75. https://doi.org/10.1038/ng0496-368



(Figure 3 from Speicher et al.[43]. © publisher)

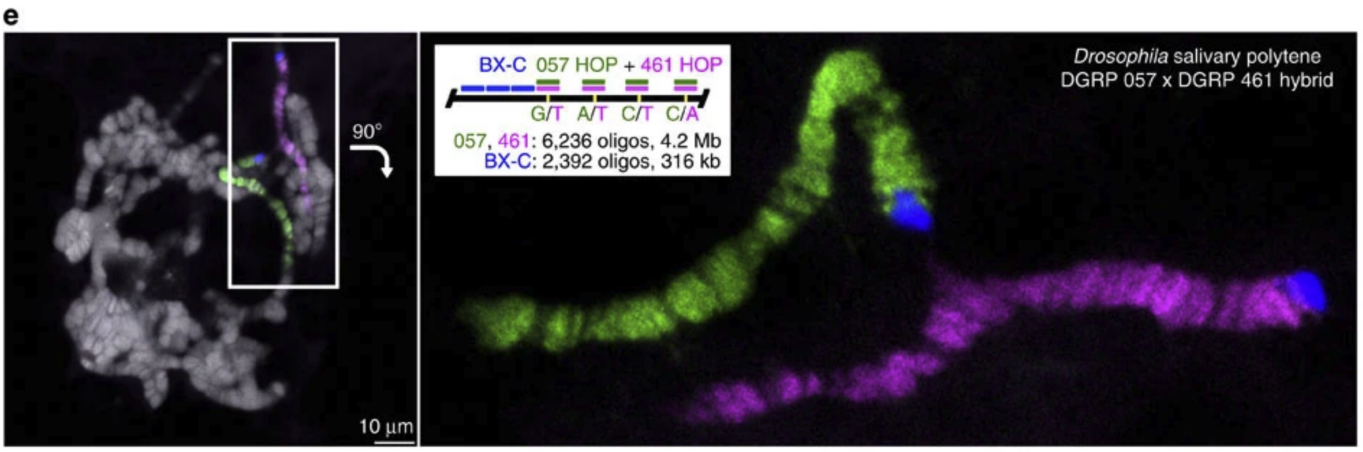
FISH is fairly fast, taking in the ballpark of an hour of bench work. The hybridization step seems to take around half a day (in an incubator)[44].

Beliveau et al. (2015) created a library of probes that were SNP-specific for different homologous chromosomes in a hybrid Drosophila line[45]. Treating denatured genomes with these probes produced pictures like this:

[43]Speicher, Michael R., Stephen Gwyn Ballard, and David C. Ward. 'Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH'. Nature Genetics 12, no. 4 (1996): 368–75. https://doi.org/10.1038/ng0496-368

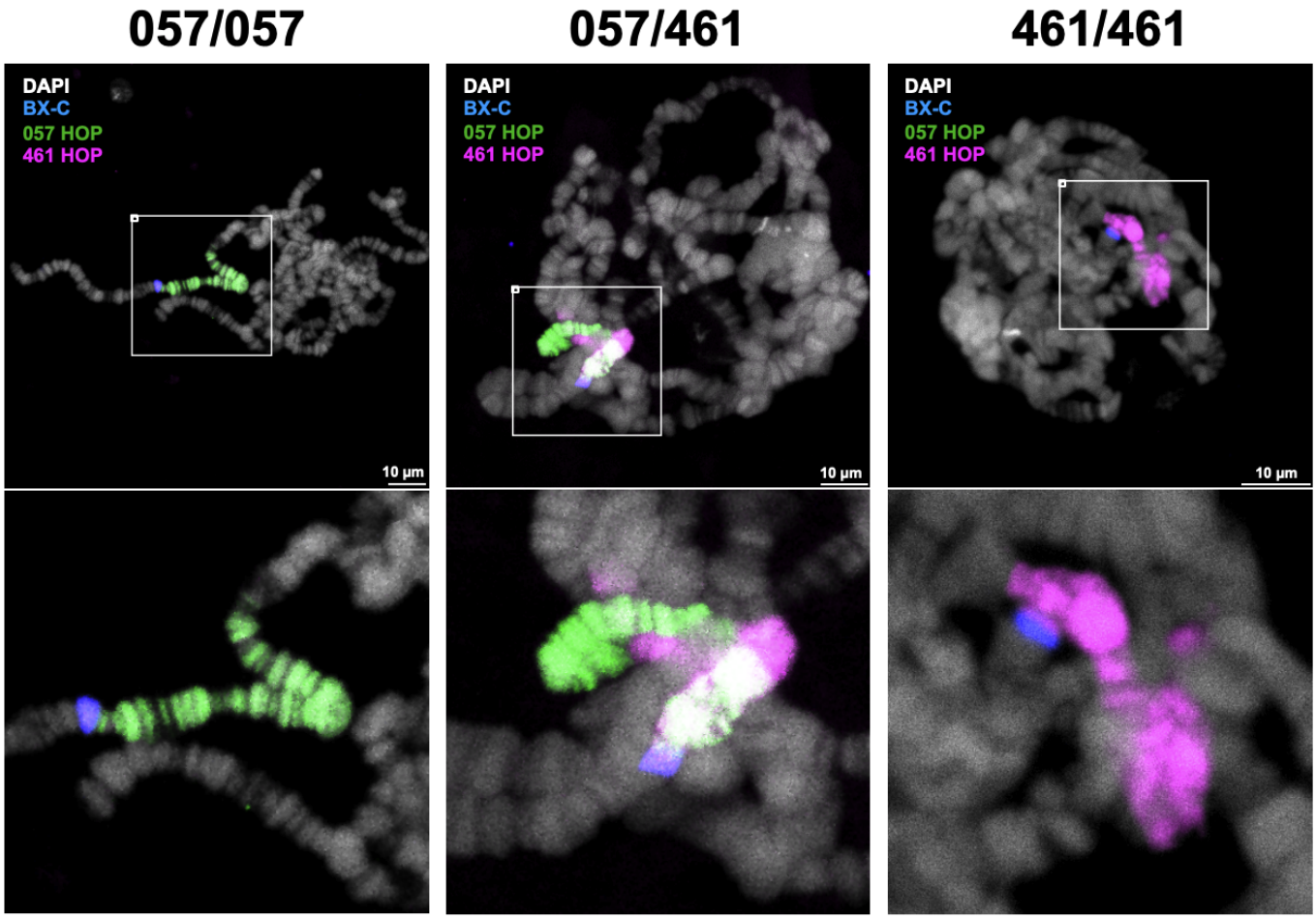
[44]Garimberti, Elisa, and Sabrina Tosi. 'Fluorescence in Situ Hybridization (FISH), Basic Principles and Methodology'. In Fluorescence in Situ Hybridization (FISH): Protocols and Applications, edited by Joanna M. Bridger and Emanuela V. Volpi. Humana Press, 2010. https://doi.org/10.1007/978-1-60761-789-1_1

[45]Beliveau, Brian J., Alistair N. Boettiger, Maier S. Avendaño, et al. "Single-Molecule Super-Resolution Imaging of Chromosomes and in Situ Haplotype Visualization Using Oligopaint FISH Probes." Nature Communications 6, no. 1 (2015): 7147. https://doi.org/10.1038/ncomms8147.



(Figure 4e from Beliveau et al.(2015). © publisher)

The two colored segments come from two homologous chromosomes (and they share the common blue stain); the green segment comes from one Drosophila line, the magenta comes from another line. The two parent lines give either only green or only magenta, given treatment with the same labels:



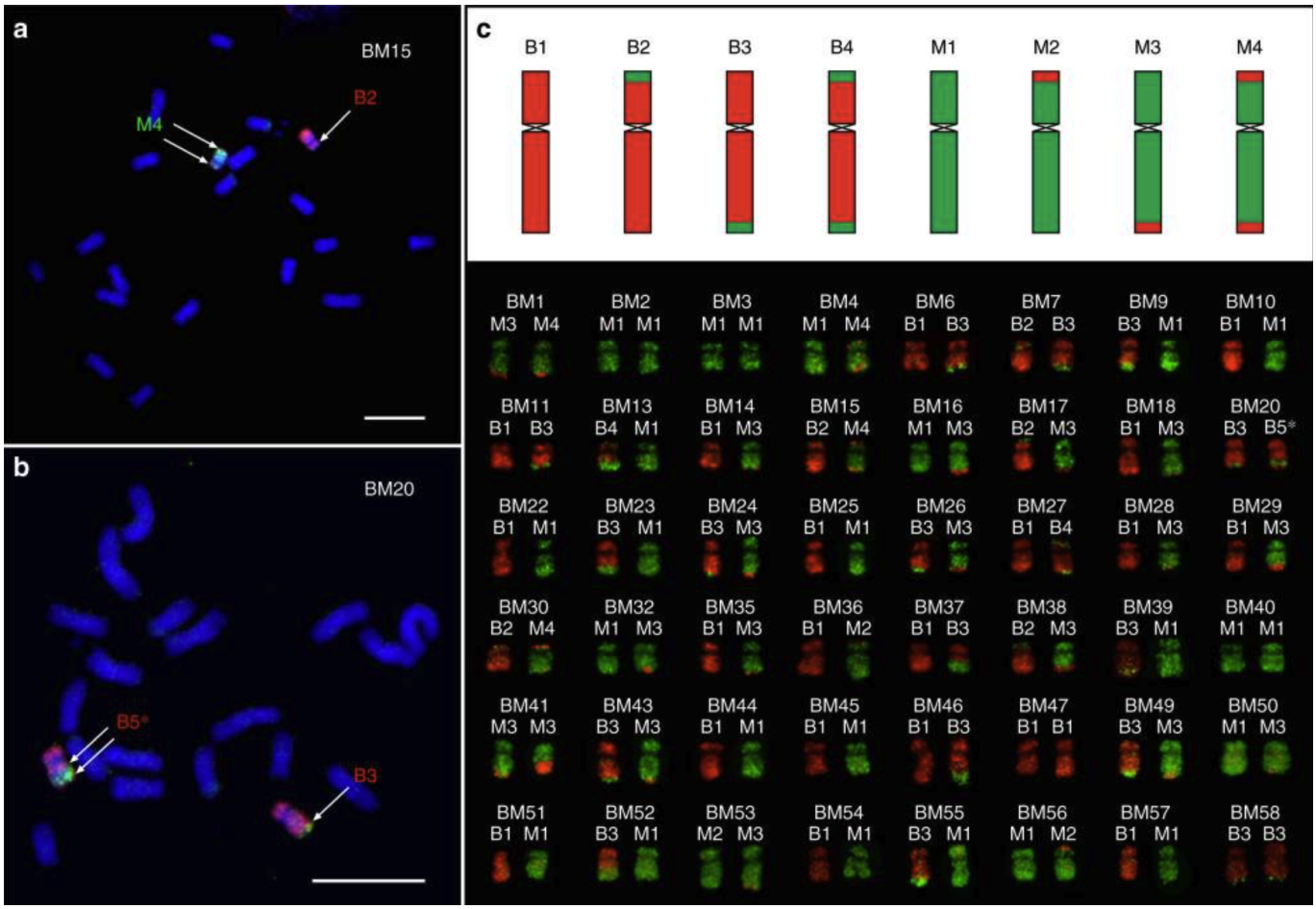
Supplementary Figure 19 HOPs FISH on spread Drosophila polytene chromosomes. FISH was performed on polytene chromosomes isolated from the salivary glands of 057/057 homozygotes (right), 057/461 hybrids (center), and 461/461 homozygotes (left). In all cases, the BX-C probe set (blue) was hybridized to the region adjacent to the 057 (green) and 461 (magenta) 89E-93C HOPs. DNA was stained with DAPI (grey). Bottom panels: zoomed-in images of boxed regions in the corresponding top panels. Images were captured on a laser scanning confocal microscope and are single Z slices. Note that because Drosophila pairs its homologous chromosomes in somatic cells, only one swath of signal is expected per locus targeted, with separation being occasionally observed (e.g. in the center panels).

(Figure S19 from Beliveau et al.(2015). © publisher)

Gametes of the F_1 hybrid (or to an F_2 hybrid, i.e. the grandchild of the two original lines) would have recombined chromosomes with 1—3 segments from the maternal line, and 1—3 segments from the paternal line. Logically, if this same protocol were applied to those chromosomes, we would see which segments were

transmitted, i.e. we could identify homologs.

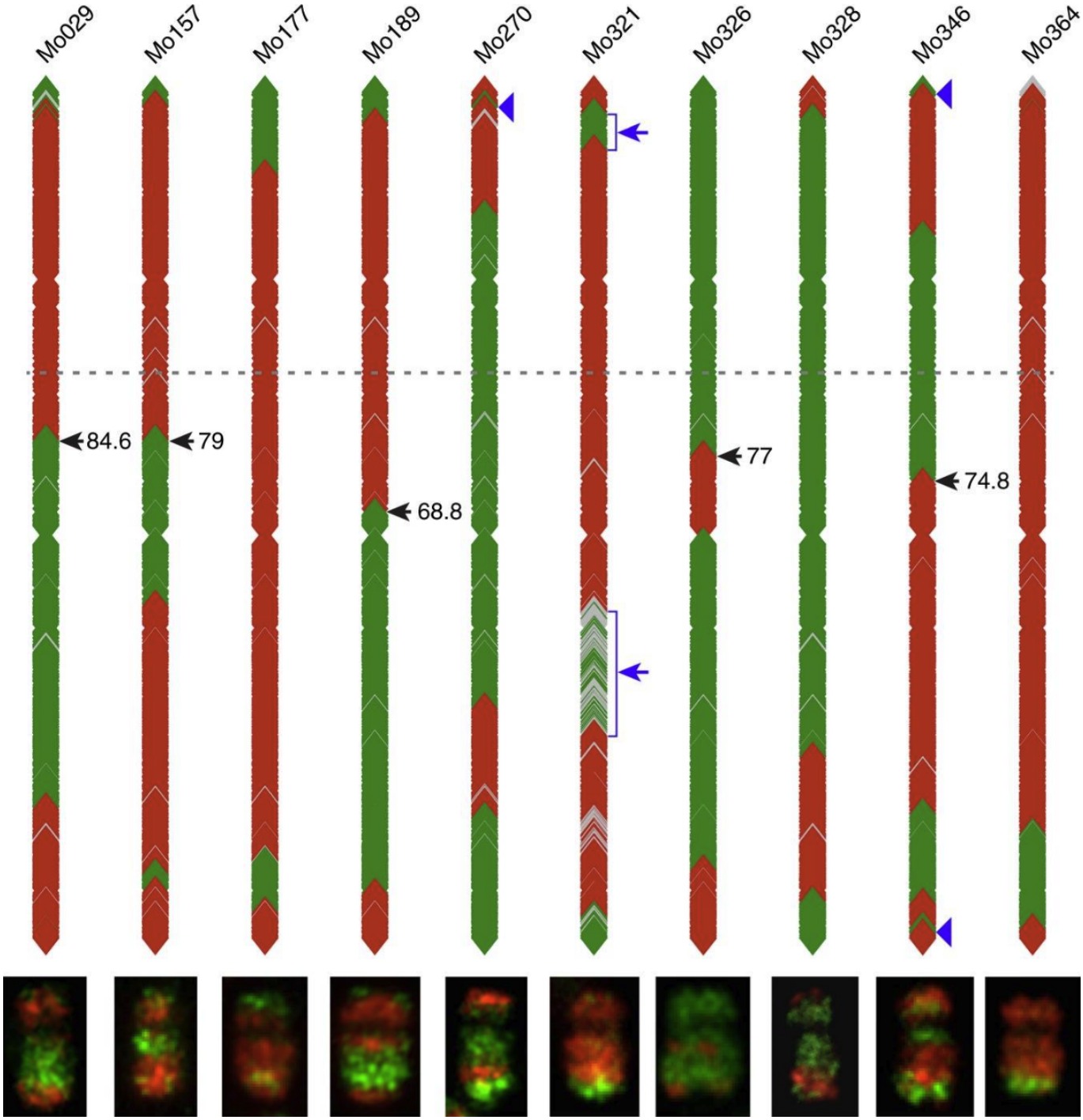
Vale Martins et al. (2019) did that very experiment with plants[46]. The F_2 generation of hybrids between two (inbred, homozygous) lines had chromosomes that showed single crossovers:



(Figure 3 from Vale Martins et al.(2019). © publisher)

They then intermated these plants for four more generations, and applied their FISH protocol to these F_6 plants:

[46]Vale Martins, Lívia do, Fan Yu, Hainan Zhao, et al. "Meiotic Crossovers Characterized by Haplotype-Specific Chromosome Painting in Maize." Nature Communications 10 (October 2019): 4604. https://doi.org/10.1038/s41467-019-12646-z.



(Figure 6 from Vale Martins et al.(2019). © publisher)

These are nice results, showing reasonably precise homolog identification.

CRISPR-dCas9 systems

It's possible to label specific sequences with fluorescent CRISPR-dCas9 systems. Like the familiar editing systems, dCas9 binds to a specific DNA segment corresponding to a guide RNA; but the "d" stands for "dead", meaning that it doesn't cut the DNA. It just binds and sticks (locally denaturing the DNA and hybridizing the guide RNA to a strand of DNA).

Deng et al. (2015) used this method to label chromatin at specific sequences[47]. They claim that this method is generally comparable to FISH, but is faster (less than an hour) and does not require DNA denaturation. They use a milder acid fixation (which may still damage chromosome structure and/or epigenomic state).

A CRISPR system tends to be roughly 200 kDa. (That's kilodaltons; one Dalton is about the weight of a proton or neutron. An amino acid is on average around 100 Daltons; a DNA base pair is around 600 Daltons.) That's more massive than a PRINS probe, at under 30 kDa; but I think it's less than a usual

[47]Deng, Wulan, Xinghua Shi, Robert Tjian, Timothée Lionnet, and Robert H. Singer. 'CASFISH: CRISPR/Cas9-Mediated in Situ Labeling of Genomic Loci in Fixed Cells'. Proceedings of the National Academy of Sciences 112, no. 38 (2015): 11870–75. https://doi.org/10.1073/pnas.1515692112

FISH probe size of many hundreds or more base pairs, which is hundreds or thousands of kDa. This might imply that a CRISPR system is better at accessing DNA in tightly packed chromatin, e.g. sperm chromatin, compared to FISH.

CRISPR-Cas9 in general requires more work than oligo-FISH, because you have to find target sequences positioned appropriately relative to PAM sequences.

On the other hand, I think a CRISPR-dCas9 system would tend to only unzip DNA at targeted locations. Further, a CRISPR-dCas9 system might require significantly less destructive treatment of the chromatin before binding, compared to FISH—I'm not sure.

There's been at least some more research on CRISPR-dCas9 systems for painting chromosomes[48][49] by number[50] and perhaps by allele / homolog[51]. But, I haven't evaluated this literature.

Other hybridization methods

- *FISH variants.*
  - Two sources of damage to chromatin in the FISH protocol are formamide treatment[52] and genome-wide denaturation by heating. There has been work on several variants of FISH, which avoid needing one or both of those steps, and which may have other advantages such as being faster.
  - See for example RASER-FISH[53], oligo-FISH[54], and GOLD FISH[55].
- *Primed in situ labeling (PRINS).*
  - PRINS is similar to FISH, but it uses unlabeled primers (short DNA strands, <50 nucleotides). DNA polymerase is added to synthesize continuations of the primers (only the hybridized ones) using some fluorophore nucleotides.
  - PRINS is much less widely used compared to FISH.
  - But, PRINS might be faster. For the labeling process itself (leaving aside the initial sample preparation), PRINS might take less than half an hour[56].
  - Also, the primers are quite small compared to FISH, which I think uses probes that tend to be 1000 bases or longer. Mozdarani and Pellestor (2004) state: "[T]his small size greatly facilitates their accessibility to their genomic target sequences. This is particularly significant in cells with highly condensed nuclei, such as spermatozoa."[57].
- *(Array) comparative genomic hybridization.*
  - I think this method is slower than other hybridization methods (it measures DNA at a much higher resolution than just chromosome numbering). It's completely destructive. But, possibly it could be made faster, if the normal protocol can be sped up in a way that sacrifices some resolution while still accurately identifying chromosome number.

[48]Hong, Yu, Guangqing Lu, Jinzhi Duan, Wenjing Liu, and Yu Zhang. 'Comparison and Optimization of CRISPR/dCas9/gRNA Genome-Labeling Systems for Live Cell Imaging'. Genome Biology 19, no. 1 (2018): 39. https://doi.org/10.1186/s13059-018-1413-5

[49]Thuma, Jenna, Yu-Chieh Chung, and Li-Chun Tu. 'Advances and Challenges in CRISPR-Based Real-Time Imaging of Dynamic Genome Organization'. Frontiers in Molecular Biosciences 10 (March 2023). https://doi.org/10.3389/fmolb.2023.1173545

[50]Zhou, Yuexin, Ping Wang, Feng Tian, et al. 'Painting a Specific Chromosome with CRISPR/Cas9 for Live-Cell Imaging'. Cell Research 27, no. 2 (2017): 298–301. https://doi.org/10.1038/cr.2017.9

[51]Maass, Philipp G., A. Rasim Barutcu, David M. Shechner, Catherine L. Weiner, Marta Melé, and John L. Rinn. "Spatiotemporal Allele Organization by Allele-Specific CRISPR Live-Cell Imaging: SNP-CLING." Nature Structural & Molecular Biology 25, no. 2 (2018): 176–84. https://doi.org/10.1038/s41594-017-0015-3.

[52]Shim, Anne R., Jane Frederick, Emily M. Pujadas, et al. 'Formamide Denaturation of Double-Stranded DNA for Fluorescence in Situ Hybridization (FISH) Distorts Nanoscale Chromatin Structure'. PLOS ONE 19, no. 5 (2024): e0301000. https://doi.org/10.1371/journal.pone.0301000

[53]Brown, Jill M., Sara De Ornellas, Eva Parisi, Lothar Schermelleh, and Veronica J. Buckle. 'RASER-FISH: Non-Denaturing Fluorescence in Situ Hybridization for Preservation of Three-Dimensional Interphase Chromatin Structure'. Nature Protocols 17, no. 5 (2022): 1306–31. https://doi.org/10.1038/s41596-022-00685-8

[54]Harun, Arrashid, Hui Liu, Shipeng Song, et al. 'Oligonucleotide Fluorescence In Situ Hybridization: An Efficient Chromosome Painting Method in Plants'. Plants 12, no. 15 (2023): 2816. https://doi.org/10.3390/plants12152816

[55]Wang, Yanbo, Wayne Taylor Cottle, Haobo Wang, et al. 'Genome Oligopaint via Local Denaturation Fluorescence in Situ Hybridization'. Molecular Cell 81, no. 7 (2021): 1566-1577.e8. https://doi.org/10.1016/j.molcel.2021.02.011

[56]Hindkjaer, Johnny, Lars Bolund, and Steen Kølvraa. 'Primed in Situ Labeling'. In Methods in Cell Biology, vol. 64. Cytometry: Part B. Academic Press, 2001. https://doi.org/10.1016/S0091-679X(01)64006-8

[57]Mozdarani, Hossein, and Franck Pellestor. 'The Primed In Situ (PRINS) Technique: An Alternative Approach for Preimplantation Chromosomal Diagnosis'. Iranian Journal of Biotechnology 2, no. 3 (2004): 149–57. https://www.ijbiotech.com/article_6915_59f6c82e1269ed50bf8c13469895396a.pdf

Standard genotyping

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Low-cover DNA sequencing		~$100?	a day or two?			1				
SNP array genotyping		~$100?	a few days?			1				

There are standard methods for measuring DNA that are used very widely:

- *DNA sequencing.*
  - This includes next-generation, Oxford nanopore, and SMRT (PacBio) sequencing.
  - Neither high-coverage nor long-read sequencing is needed for homolog or number identification. Bulk sequencing should be cheap, and should take a day or two or something.
- *SNP genotyping.*
  - This should also be cheap and fast. The arrays can be small.

(See also RNA sequencing, as well as the subsections "Sequencing post-meiotic RNA" and "Sequencing a cell culture" on indirect identification methods.)

These methods work via a close interaction with DNA molecules (hybridization and/or synthesis), and they require a substantial quantity of DNA, not just a single molecule. So a polymerase-based method (PCR, MDA) is used to amplify the DNA (i.e. make copies of fragments of the DNA), and then the fragments are measured.

There are several amplification methods, including PCR (polymerase chain reaction, wiki) and others[58]. Some methods, such as degenerate oligonucleotide-primed PCR[59], are able to quickly amplify tiny amounts of DNA. MDA (multiple displacement amplification), for example, has been used to amplify DNA from single cells[60] and even single chromosomes (very large wheat chromosomes—$O(10^9)$ bp)[61].

DNA polymerase methods work by denaturing the DNA (unzipping the double helix into two strands) so that the DNA primers and DNA polymerase can bind to the single-stranded DNA. The original DNA molecule is damaged by the steps of deproteinizing (e.g. with proteinase K) and then denaturing (usually by heating, sometimes with enzymes).

Therefore, these methods are almost always destructive.

However, they are generally fast, cheap, accurate, and well-developed (meaning that equipment and expertise, and/or platform services, are available). So they are good candidates for complementary identification. One can do PCR and then do low-coverage sequencing; that should be more than enough to confidently identify the chromosome number (and homolog of that chromosome).

In theory, amplification doesn't necessarily require destroying the DNA. This is similar to other methods that involve hybridization or otherwise interacting with unzipped DNA. If it is the case that…

- …a small amount of interaction with the naked DNA (single-stranded, or at least deproteinized double-stranded) is sufficient for chromosome identification,
- …and if that small amount of interaction can be enabled with only a small amount of denaturing,
- …and if the whole chromosome can be handled gently throughout the process and separated away from the products/debris involved in the identification method,

[58] Huang, Lei, Fei Ma, Alec Chapman, Sijia Lu, and Xiaoliang Sunney Xie. 'Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications'. Annual Review of Genomics and Human Genetics 16, no. Volume 16, 2015 (2015): 79–102. https://doi.org/10.1146/annurev-genom-090413-025352

[59] Telenius, Håkan, Nigel P. Carter, Charlotte E. Bebb, Magnus Nordenskjöld, Bruce A. J. Ponder, and Alan Tunnacliffe. 'Degenerate Oligonucleotide-Primed PCR: General Amplification of Target DNA by a Single Degenerate Primer'. Genomics 13, no. 3 (1992): 718–25. https://doi.org/10.1016/0888-7543(92)90147-K

[60] Lasken, Roger S. 'Single-Cell Genomic Sequencing Using Multiple Displacement Amplification'. Current Opinion in Microbiology, Antimicrobials/Genomics, vol. 10, no. 5 (2007): 510–16. https://doi.org/10.1016/j.mib.2007.08.005

[61] Cápal, Petr, Nicolas Blavet, Jan Vrána, Marie Kubaláková, and Jaroslav Doležel. 'Multiple Displacement Amplification of the DNA from Single Flow–Sorted Plant Chromosome'. The Plant Journal 84, no. 4 (2015): 838–44. https://doi.org/10.1111/tpj.13035

…then in theory one could directly identify chromosomes this way, and then use them in a functional genome. But this hasn't been attempted as far as I know, and might be very difficult.

Bulk methods

A bulk method for chromosome identification is one that handles many chromosomes at once without tracking which chromosomes came from the same original cell as each other. You might lyse a bunch of cells, producing buffer fluid with many chromosomes in it, all mixed together. Then you sort them, all at once, in bulk. The desired output is some fraction of the slurry that contains mostly chromosomes of one number.

Bulk methods in general

Bulk methods have some problems:

- A bulk method for chromosome identification is one that destroys origin indexing information.
  - It therefore can't be used for indirect identification. (See the section "Indirect identification requires origin indexing" above.) For this reason, a bulk method tends to be poor for confidence. (See the section "Setwise calls vs. uncertain calls" above.)
  - One exception is if the bulk method gets high-confidence calls for all or almost all single chromosome numbers.
  - As an additional consequence, bulk methods do not combine with setwise homolog identification. If you could magically learn the homolog (allele) content of a single sperm cell, but your only way to identify the number of individual chromosomes was a bulk method, then you could not do homolog-wise chromosome selection.
- Another issue is that a bulk method would tend to handle chromosomes in an uncontrolled way, exposing them to shear forces and hence DNA damage.

The appeal of bulk methods is that they deal with large numbers. So, they tend to be very very cost efficient, maybe by several orders of magnitude. However, in a context where we can't culture cells to verify euploidy, what we want is high confidence in a single euploid cell, not many cells some of which are euploid. These problems make bulk methods probably not workable, at least as part of isolating-ensembling methods for chromosome selection.

This points at another exception to the rule that bulk methods aren't workable. If we're using the bulk identification method within a chromosome selection protocol where we can verify results, e.g. by culturing and sequencing, then it could be very useful. For example, one could use centrifugation to get a slurry that contains many chromosomes, which are mostly number 1; then produce many microcells (somehow); then apply the second half of MMCT to obtain many cells that receive the exogenous chromosome number 1. Compared to getting microcells with a random chromosome, this is a significant quantitative improvement.

Centrifugation

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Centrifuga		??	minutes or hours?			ambiguous sets up to ~4 by mass	bulk method; not confident	high shear forces?	?	?

A standard tool for studying the contents of cells is centrifugation. You lyse a sample of cells and then spin the resulting suspension really fast, so that particles form sedimentary layers depending on shape and mass. Centrifugation can be used to pull chromosomes out of such a suspension[62].

A hypothetically possible method for chromosome sorting is sedimentation using a centrifuge. You have a chromosome suspension with dispersed individual chromosomes. Then you spin it extremely fast in a centrifuge; depending on shape and mass, chromosomes of different number settle in different bands.

[62] Gasser, S. M., T. Laroche, J. Falquet, E. Boy de la Tour, and U. K. Laemmli. 'Metaphase Chromosome Structure: Involvement of Topoisomerase II'. Journal of Molecular Biology 188, no. 4 (1986): 613–29. https://doi.org/10.1016/S0022-2836(86)80010-9

This method can work at least partially. Stubblefield et al. (1975) were able to somewhat sort hamster chromosomes by centrifugation[63]. The separation is far from complete, but substantial:

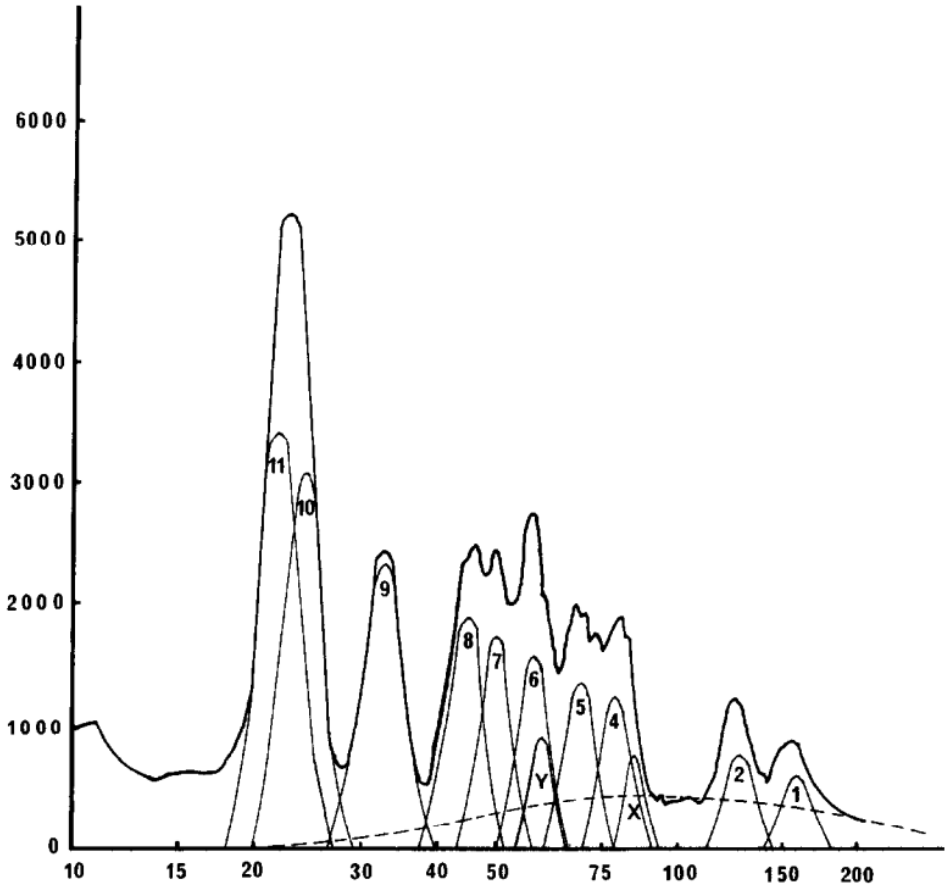


Fig. 2. Abscissa: channel no. (proportional to DNA content); *ordinate:* particles/channel.

FMF pattern produced by chromosomes isolated from Don-6 cells. The overall pattern is resolvable into a series of overlapping peaks, as shown. The channel numbers are plotted exponentially so that the peak widths remain similar; the peak heights diminish to the right, since the peak areas are identical and the peak widths are actually greater. The peaks are numbered according to the chromosomes (see fig. 1) assigned to each peak. The broad dotted peak is thought to be the result of random aggregation of chromosomes.

(Figure 2 from Stubblefield et al. (1975). © publisher)

A similar experiment by Collard et al. (1984) showed substantial separation of human chromosomes[64]. They used a centrifuge exerting a force of about $50g$. This figure (only part of the full figure) shows the scattering spectra of samples from different sediment bands:

[63]Stubblefield, E., S. Cram, and L. Deaven. 'Flow Microfluorometric Analysis of Isolated Chinese Hamster Chromosomes'. Experimental Cell Research 94, no. 2 (1975): 464–68. https://doi.org/10.1016/0014-4827(75)90519-4

[64]Collard, J. G., E. Philippus, A. Tulp, R. V. Lebo, and J. W. Gray. 'Separation and Analysis of Human Chromosomes by Combined Velocity Sedimentation and Flow Sorting Applying Single- and Dual-Laser Flow Cytometry'. Cytometry 5, no. 1 (1984): 9–19. https://doi.org/10.1002/cyto.990050104

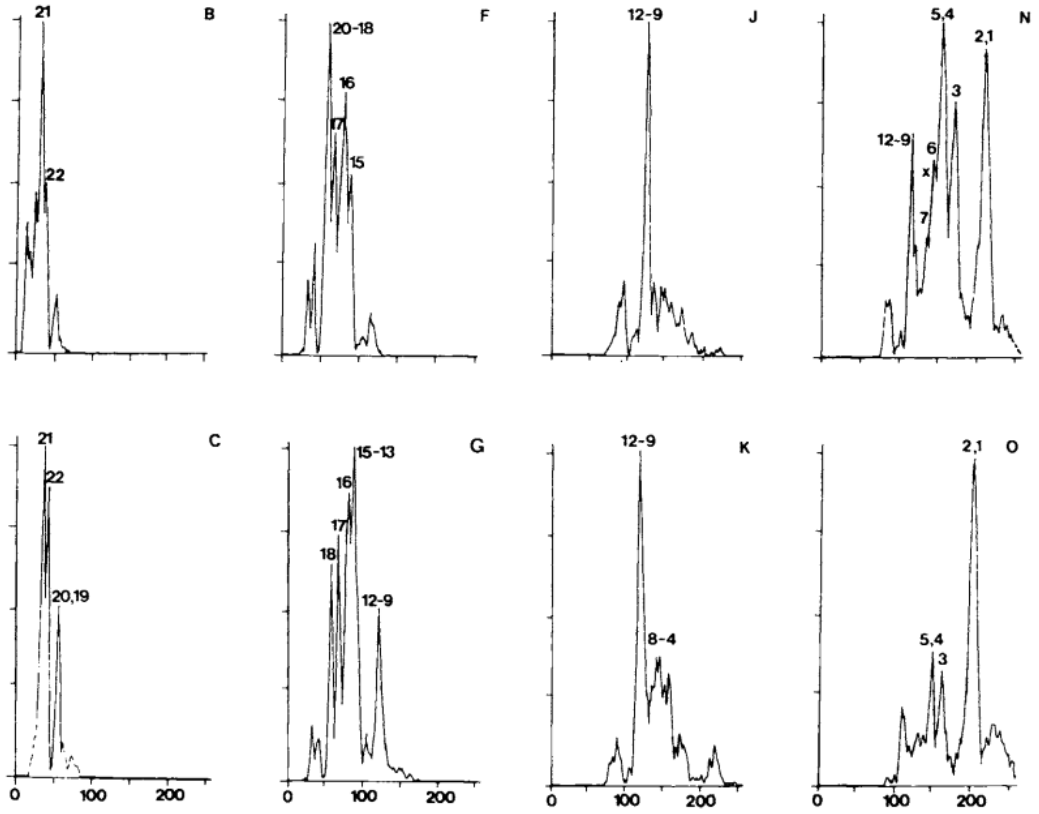


FIG. 4. Flow cytometric histograms of the human chromosomes separated by velocity sedimentation at 52g and stained with the mixture mithramycin and ethidium bromide. Abscissa, channel number x, the relative amount of DNA per chromosome, ordinate, distribution f(x), the relative number of chromosomes per channel. Panel HFL) Unfractionated chromosomes analysed before layering on the gradient. Panels A–O) The chromosomal composition of the different 4-ml fractions after 52g sedimentation. Numbers 1–22 and X,Y represent the position of the different human chromosomes in the histograms.

(Figure 4 from Collard et al. (1984). © publisher)

But again, there remains a lot of ambiguity. In theory centrifugation could separate all human chromosomes, maybe by spinning them faster (ultracentrifugation). This has been speculated about, e.g. in Noll and Noll (1989)[65]. But I haven't seen this done in the literature.

Further, there's some reason to think some ambiguities won't go away. Van Dyke et al. (1986) give these numbers for metaphase chromosomes taken from five human males (the lengths are relative, so the autosomes (1—22) sum to 100)[66]:

[65]Noll, Hans, and Markus Noll. '[5] Sucrose Gradient Techniques and Applications to Nucleosome Structure'. In Methods in Enzymology, edited by Paul M. Wassarman and Roger D. Kornberg, vol. 170. Nucleosomes. Academic Press, 1989. https://doi.org/10.1016/0076-6879(89)70043-4

[66]Van Dyke, D. L., Maria J. Worsham, L. J. Fisher, and L. Weiss. 'The Centromere Index and Relative Length of Human High-Resolution G-Banded Chromosomes'. Human Genetics 73, no. 2 (1986): 130–32. https://doi.org/10.1007/BF00291602

Table 1. Relative length and centromere index of human chromosomes. The relative lengths and centromere indexes in midmetaphase and high resolution chromosome preparations are compared. SD, standard deviation. Any significant difference between midmetaphase and high-resolution at overall alpha-level of 0.05 is indicated by an *asterisk*

Chromosome number	Relative length				Centromere index			
	Mid-metaphase		High-resolution		Mid-metaphase		High-resolution	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	9.00	0.42	9.06	0.17	51.20	1.55	53.83	0.98
2	8.37	0.17	8.35	0.09	39.42	0.45	37.72	1.40
3	6.96	0.20	6.81	0.14	46.36	0.95	46.73	3.53
4	6.44	0.19	6.22	0.13	27.44	1.38	25.37	3.83
5	6.16	0.14	6.02	0.05	26.59	0.95	26.53	2.55
6	5.95	0.18	5.90	0.10	35.69	3.49	37.77	1.05
7	5.38	0.18	5.53	0.05	38.61	0.61	35.59	0.39*
8	4.77	0.16	4.83	0.12	34.16	0.91	32.37	1.31
9	4.76	0.14	4.71	0.12	33.81	1.73	32.69	1.77
10	4.69	0.25	4.69	0.14	33.57	1.06	30.56	1.33
11	4.63	0.07	4.58	0.11	37.36	2.04	37.86	1.64
12	4.70	0.10	4.74	0.18	29.05	0.85	28.14	1.63
13	3.41	0.04	3.36	0.10	10.74	0.91	9.50	0.40
14	3.32	0.07	3.34	0.09	11.61	0.32	9.40	0.47
15	3.42	0.20	3.31	0.12	12.25	1.00	9.98	0.51
16	3.27	0.11	3.47	0.04	44.95	0.93	46.63	0.61
17	3.21	0.09	3.42	0.13	32.52	2.05	27.79	2.67
18	2.75	0.11	2.59	0.03	28.26	0.86	24.64	1.55
19	2.63	0.07	2.78	0.10	45.55	2.18	48.32	0.71
20	2.33	0.09	2.49	0.13	44.41	1.77	42.91	3.34
21	1.41	0.08	1.49	0.07	18.33	3.20	18.06	1.50
22	1.90	0.06	1.86	0.10	19.87	1.85	16.89	0.50
X	5.08	0.11	5.18	0.10	40.13	2.64	37.51	2.60
Y	2.02	0.30	1.72	0.22	27.84	2.07	23.49	2.87

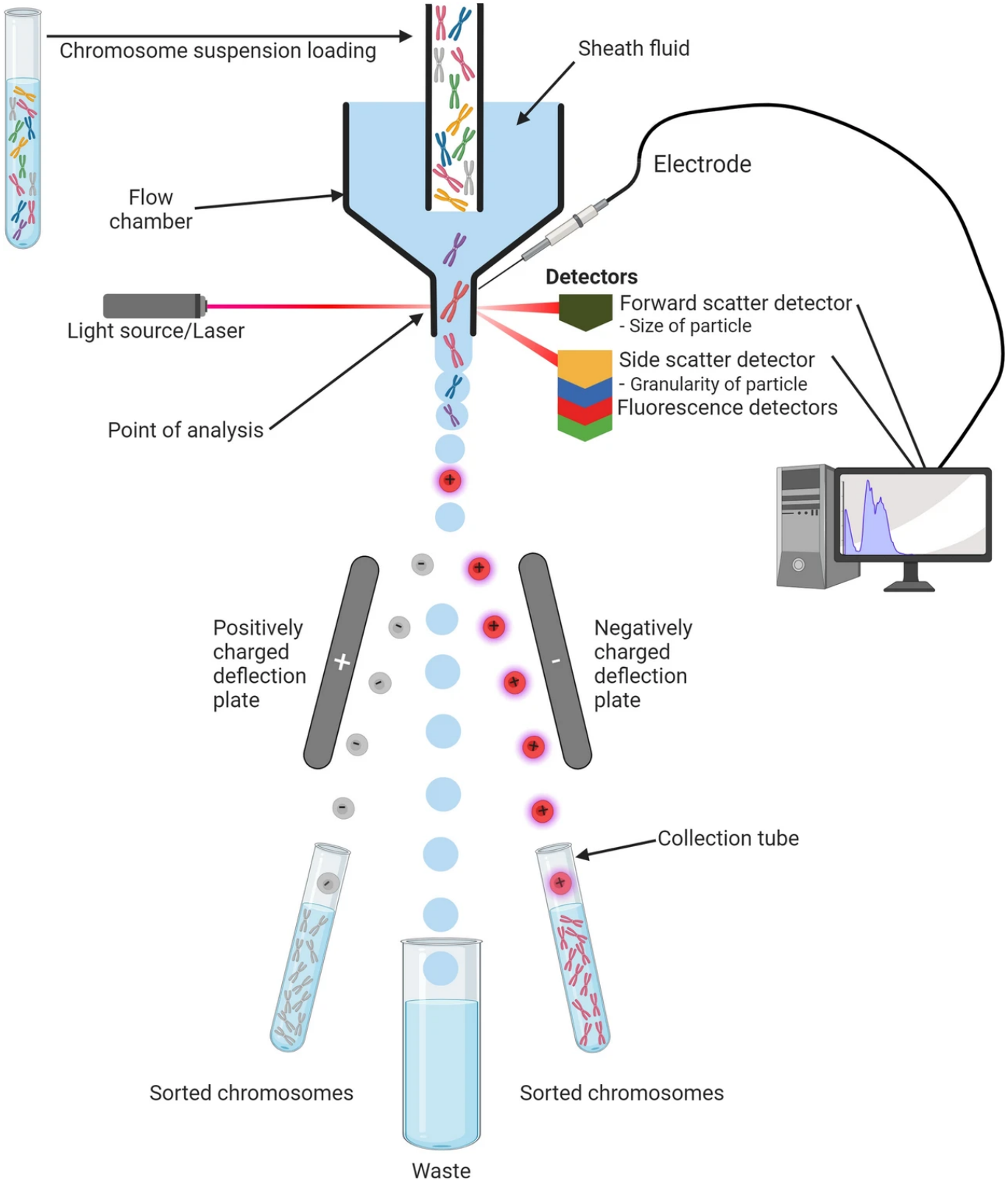
(Table 1 from Van Dyke et al. © publisher)

I don't know how much noise to expect from their measurement method. But if the SD numbers could be taken as mainly coming from actual variation, and if ultracentrifugation would sort by length, then for example chromosome numbers $\{10, 11, 12\}$ can't be confidently distinguished this way. This is a major obstacle for using centrifugation as the primary method for chromosome numbering: ambiguity between a set of four chromosomes, for example, would cut the chances of ensembling a euploid genome by a factor of 10. (See the subsection "Setwise calls vs. uncertain calls".)

Flow cytometry

A widely used method is flow cytometry. This is basically staining or FISH, except after we label the chromosomes, we shoot large numbers of labeled chromosomes really fast (1000+ per second) through the optics setup. We use electrostatic deflection to sort target from non-target chromosomes (inside droplets).

This method produces large numbers of identified chromosomes. See this illustration:



(Figure 1 from Doležel et al. (2023)[67]. © publisher)

Flow cytometry can be used to sort human chromosomes by number. For example, Langlois et al. (1982) used flow cytometry on chromosomes stained with two fluorescent dyes, and were able to classify chromosomes by number almost completely[68]:

[67]Doležel, Jaroslav, Petr Urbiš, Mahmoud Said, Sergio Lucretti, and István Molnár. 'Flow Cytometric Analysis and Sorting of Plant Chromosomes'. The Nucleus 66, no. 3 (2023): 355–69. https://doi.org/10.1007/s13237-023-00450-6

[68]Langlois, R G, L C Yu, J W Gray, and A V Carrano. 'Quantitative Karyotyping of Human Chromosomes by Dual Beam Flow Cytometry.' Proceedings of the National Academy of Sciences 79, no. 24 (1982): 7876–80. https://doi.org/10.1073/pnas.79.24.7876

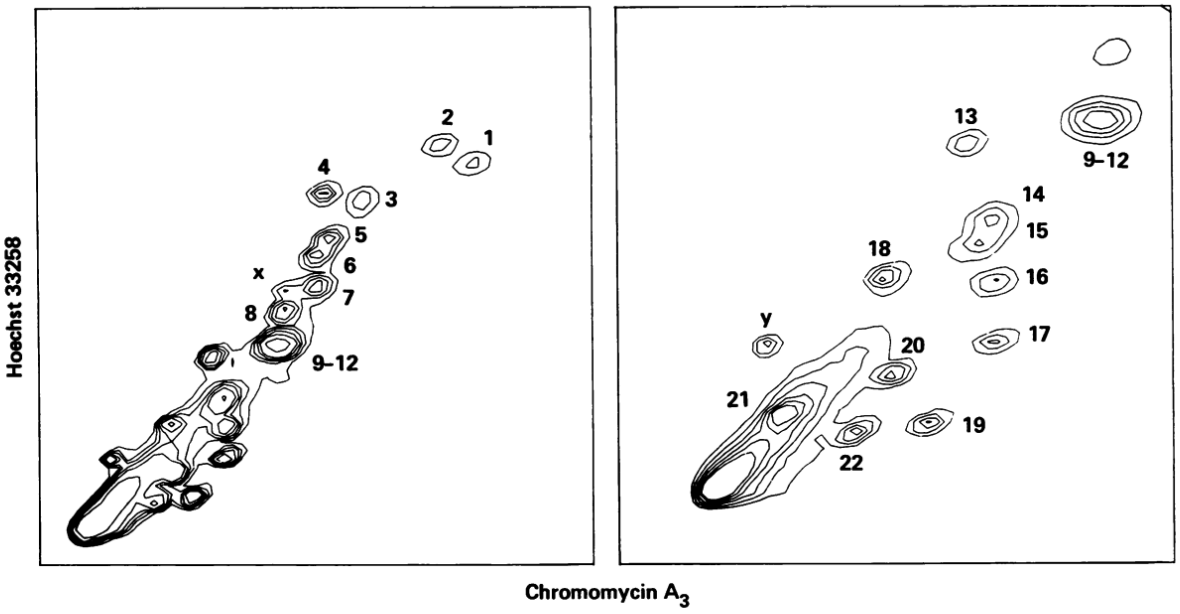


FIG. 1. (*Left*) Bivariate flow karyotype of human lymphocyte chromosomes from individual number 400. The height of each peak in the distribution is indicated by contours. (*Right*) Expanded view of chromosomes smaller than chromosome 8. Peak widths are 2–6% (standard deviation/mean).

(Figure 1 from Langlois et al. (1982). © publisher)

They say they can distinguish all chromosomes except {9,10,11,12} and {14,15}. That's fairly impressive, given the high throughput. However, since flow cytometry is a bulk method, even this is usually an unacceptable level of ambiguity. Assembling a haploid genome from chromosomes sorted this way, by randomly sampling four chromosomes and two chromosomes from those two classes, would give less than a 5% chance of a euploid genome.

Better apparatuses and different dyes have improved sorting capabilities for human chromosomes. But a 2021 review by Stanley et al. states that "it is still not possible to clearly separate all chromosomes with chromosomes 10—12 forming a cluster, due to their similarity in size and base pair composition"[69]. (That level of ambiguity, about 20% success rate for randomly sampling a correct set of 3, could be acceptable depending on context.)

In addition to the disadvantages of being a bulk method, flow cytometry also has the disadvantages of staining or FISH.

Further, the high-throughput fluidics is likely to damage the chromosomes. E.g. chromosomes could be broken by large shear forces, which cause problems in cell sorting because they hurt cell viability[70]. Doležel et al. treated chromosomes with formaldehyde to help protect them from shear forces; but formaldehyde also causes damage. Suspensions produced by flow methods tend to have lots of small fragments of chromosomes[71], which means many chromosomes are breaking, and likely there are large chromosome fragments that have had a small piece broken off (making them unacceptable for our purposes).

Electrokinesis

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Electrokin	not much applied to full chrs	??	hours?	hasn't been shown		multiple features might ID singles?	?	shear forces?	?	?

[69] Stanley, Jason, Henry Hui, Wendy Erber, Britt Clynick, and Kathy Fuller. 'Analysis of Human Chromosomes by Imaging Flow Cytometry'. Cytometry Part B: Clinical Cytometry 100, no. 5 (2021): 541–53. https://doi.org/10.1002/cyto.b.22023

[70] '4 Ways Your Droplet Sorter May Harm Your Cells | Miltenyi Biotec | USA'. Accessed 16 September 2025. https://www.miltenyibiotec.com/US-en/resources/blog/4-ways-your-droplet-sorter-may-harm-your-cells.html.

[71] Doležel, Jaroslav, Sergio Lucretti, István Molnár, Petr Cápal, and Debora Giorgi. 'Chromosome Analysis and Sorting'. Cytometry Part A 99, no. 4 (2021): 328–42. https://doi.org/10.1002/cyto.a.24324

Electrokinetic methods use electrical fields to move around particles that aren't electrically identical to the medium. Particles with different electrical properties respond differently to some applied electrical fields, so they can be sorted. Chromosomes with different size and attached proteins should have different charges and polarizabilities, so maybe they can be identified through electrokinesis.

These methods are potentially appealing because they should be non-destructive, acting at a distance. They've been used by several groups to sort small ($< O(10^6)$ base pair) DNA molecules.

However, common electrophoresis methods don't work well for large DNA molecules. Dielectrophoresis or less common variants of gel electrophoresis might work, but they don't seem to have been tried much for full human chromosomes.

Types of electrokinesis

In general, electrokinesis is motion induced by electrical forces. The two main kinds of electrokinetic methods for particle sorting are electrophoresis and dielectrophoresis.

Electrophoresis is simple enough: An electric field is applied (direct current). Charged particles experience an electrostatic force relative to an uncharged medium, so the particles are pulled around:

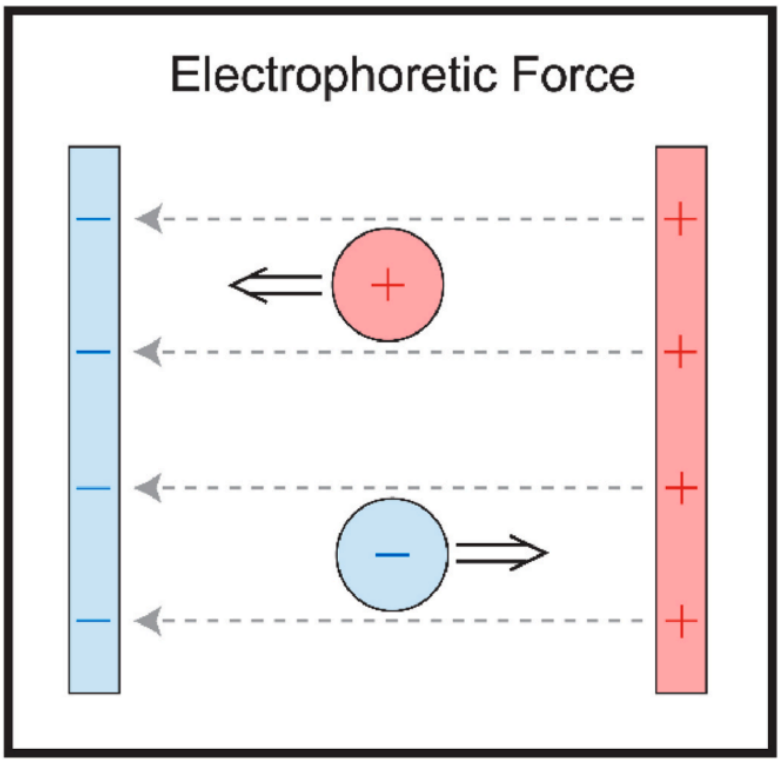
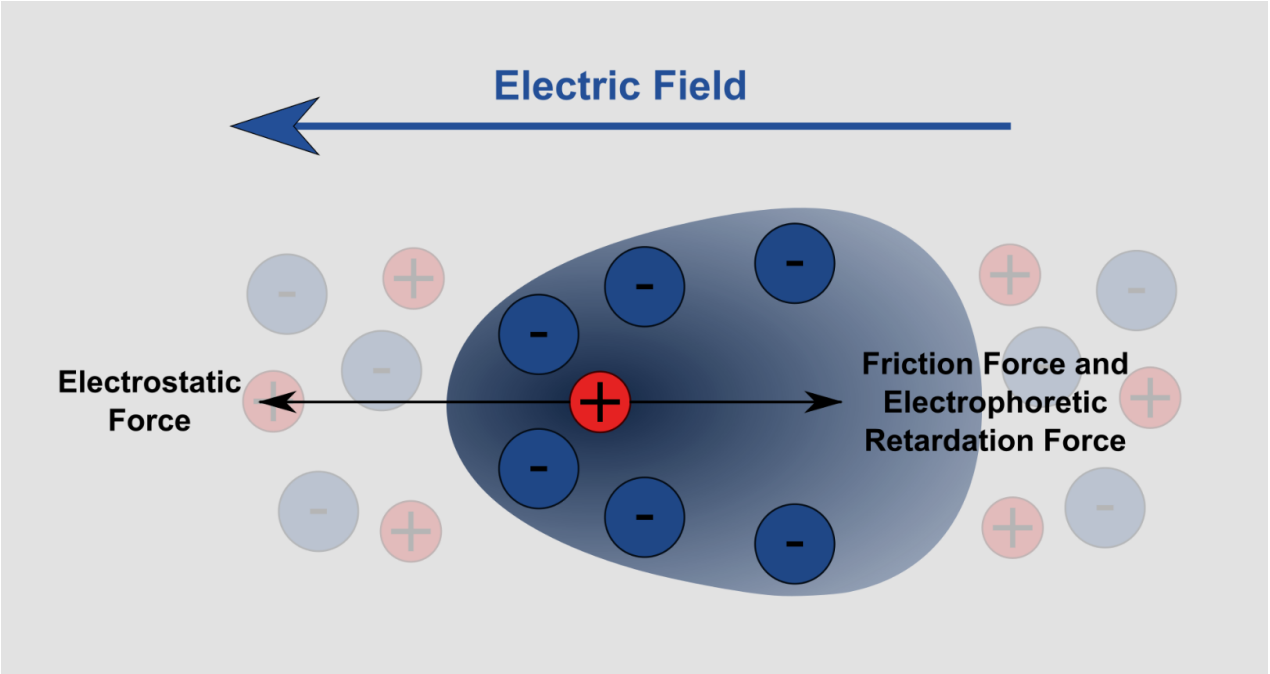


Fig. 1. Illustration of how electrophoresis interacts with charged particles and the direction the different charged particles will move in the presence of a direct current power and a uniform electric field.

(Figure 1 from Hagness et al. (2023)[72]. © publisher)

(The actual set of forces is a bit more complicated:)

[72]Hagness, Daniel E., Ying Yang, Richard D. Tilley, and J. Justin Gooding. 'The Application of an Applied Electrical Potential to Generate Electrical Fields and Forces to Enhance Affinity Biosensors'. Biosensors and Bioelectronics 238 (October 2023): 115577. https://doi.org/10.1016/j.bios.2023.115577



(Visualization by Daniele Pugliesi. © publisher)

Dielectrophoresis is when a dielectric particle is moved by a non-uniform electric field. A dielectric particle is a particle that gets polarized by being in an electric field—i.e. one side of the particle becomes positively charged and the other side becomes negatively charged. (Most particles are at least somewhat dielectric: charge can, to some small extent, slosh around within the particle when it's in an electric field.) The particle's net charge doesn't change from polarization, so if the particle was neutral, it's still neutral overall. If the electric field were constant, the forces on the polarized particle would cancel out and the particle wouldn't experience net force. However, if the electric field is non-uniform, then the positively charged side and the negatively charged side might be in electric fields of slightly different strengths. Then the whole particle experiences a net force:

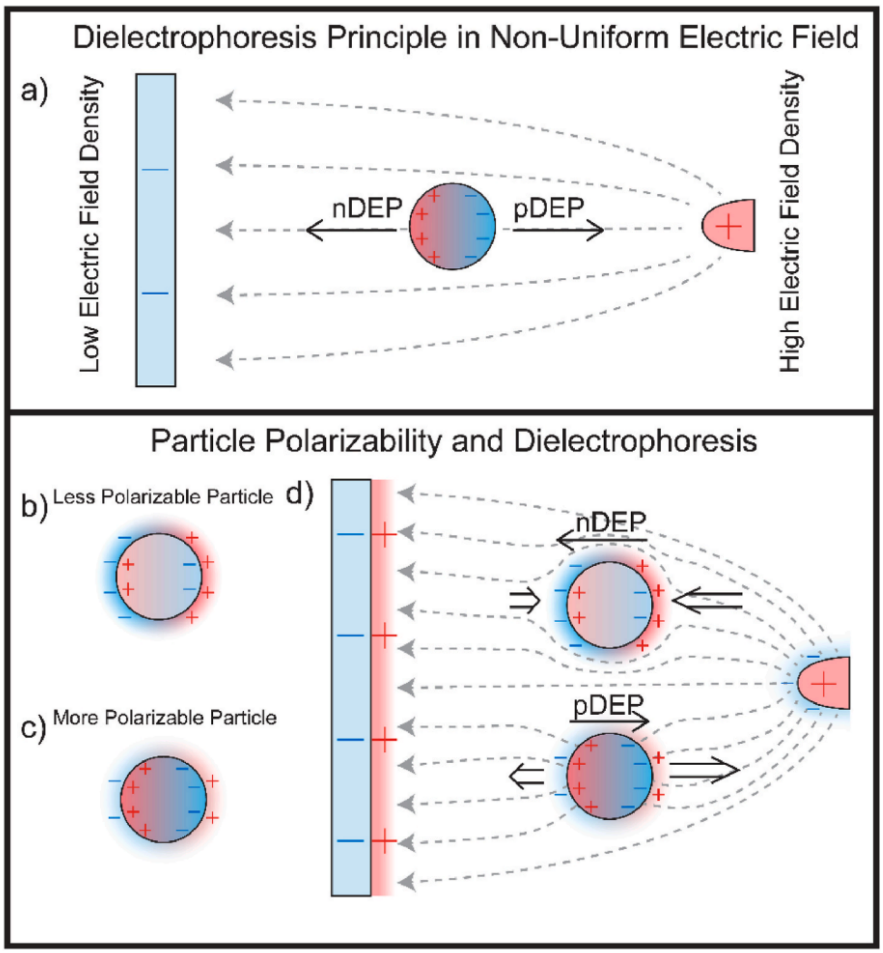


Fig. 4. a) Illustrates the difference between positive dielectrophoresis (pDEP) and negative dielectrophoresis (nDEP). A dipole particle in a non-uniform electric field will either undergo positive dielectrophoresis, which is where the particle moves towards the region of high electric field density, or negative dielectrophoresis, which is where the particle moves towards the region of low electric field density. b) schematic of a particle with a lower polarizability than the medium. c) schematic of a particle with a higher polarizability than the medium. d) Shows the different dielectrophoresis effect particle of different polarizabilities experience. For a particle that is less polarisable than the medium, it will experience negative dielectrophoresis. This is due to the greater charge build-up on the medium side of the particle-medium interface, which will result in the particle being pulled towards the region of low electric field density. For a particle that is more polarisable than the medium, it will experience positive dielectrophoresis. This is because the charge-up is the inverse of the previous case and more charge builds-up on the particle side of the particle-medium interface which will result in a pulling of the particle towards the region of high electric field density.

(Figure 4 from Hagness et al. (2023). © publisher)

Depending on the parameters of the electric field, it's possible to have electrophoretic forces but not dielectrophoretic forces: A charged particle (electrophoretic force), with a uniform electric field or with polarizability that matches the polarizability of the medium (no dielectrophoretic force). You can also have the reverse: An uncharged particle (no electrophoretic force), with dielectrophoretic force produced by a non-uniform electric field and particle polarizability that's substantially different from the polarizability of the medium.

Both of these forces can be used to move DNA around, either through a liquid medium or through gel (usually agarose or polyacrylamide). In many circumstances, differently-sized DNA molecules move at different rates. Thus, differently-sized DNA molecules can be distinguished by where they end up after electrokinesis. (Gel electrophoresis of DNA is a key element of Sanger sequencing, the dominant sequencing method for about two decades until next-generation sequencing was developed.)

(Electroosmosis and the electrothermal effect are other electrokinetic effects, but I don't know if they can be used for differentially moving suspended particles.)

Electrophoresis

The theory of gel electrophoresis has been elaborated for a few decades in the literature[73]. Jones et al. (2017) note that how much a DNA molecule moves through gel under electrophoretic forces is probably *independent* of DNA length beyond some point ($> O(10^4)$ bp), citing a theoretical prediction[74]. Empirically, long DNA molecules do indeed move through gel at similar rates. Fangman (1978) tried low concentrations of agarose gel and low voltages[75]. That helped with separation up to $O(10^5)$ bp (whereas human chromosomes are $O(10^8)$ bp), but the longer molecules still appeared to converge in mobility:

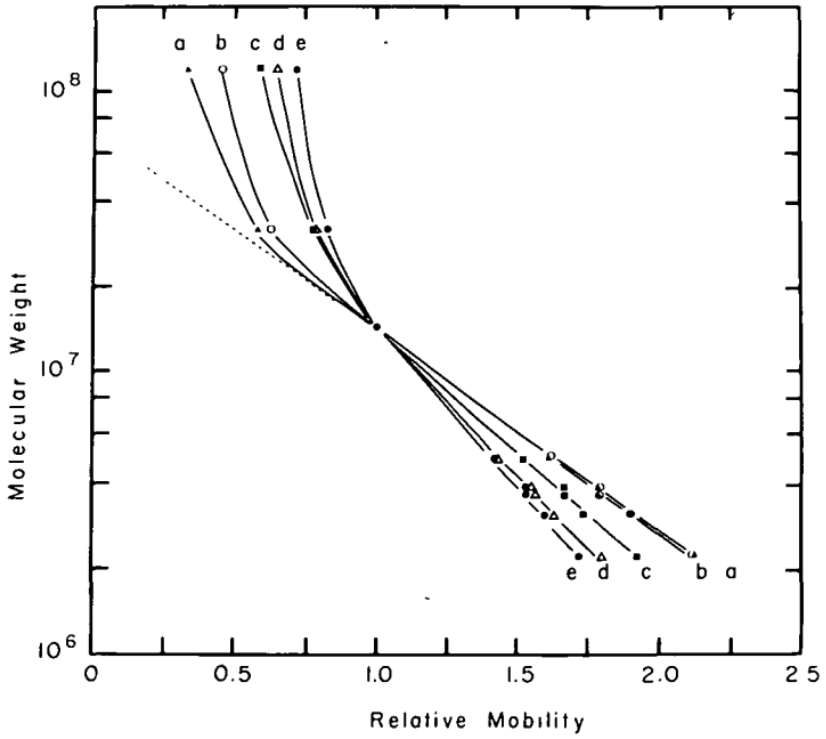


Figure 2. Effect of voltage on DNA electrophoretic mobility. The agarose concentration was 0.2%. The lines represent: (a) 0.02 V/cm, (b) 0.1 V/cm, (c) 0.5 V/cm, (d) 1.0 V/cm and (e) 2.5 V/cm. Mobilities are normalized to the mobility of Eco RI λ DNA fragment A which had the following value in each gel: (a) 0.34 mm/hr, (b) 0.64, (c) 2.2, (d) 3.9 and (e) 8.7. The dashed line is the linear extrapolation of the data for low molecular weight DNAs in experiment a.

(Figure 2 from Fangman (1978). © publisher)

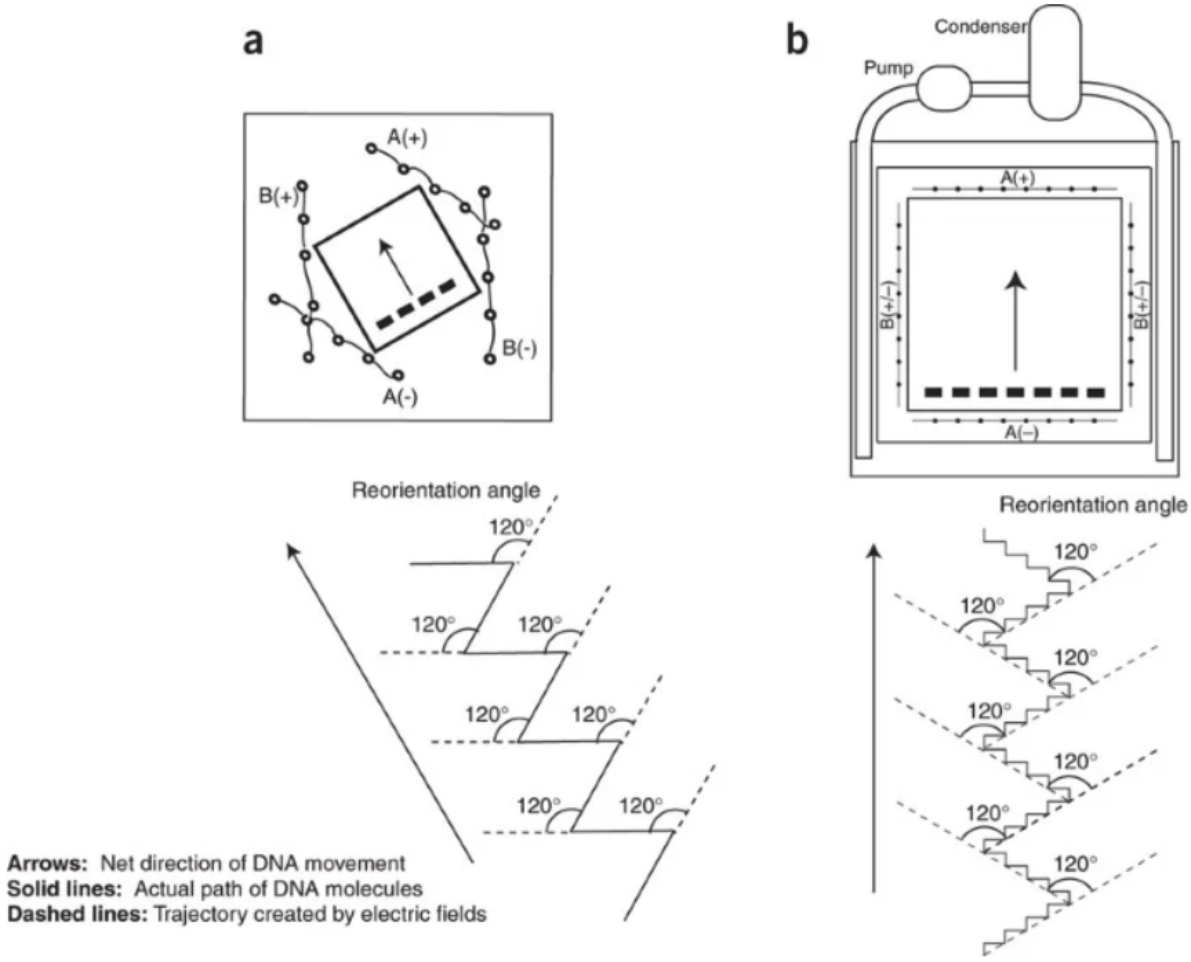
[73]Slater, Gary W. 'DNA Gel Electrophoresis: The Reptation Model(s)'. ELECTROPHORESIS 30, no. S1 (2009): S181–87. https://doi.org/10.1002/elps.200900154

[74]Duke, T. A. J., A. N. Semenov, and J. L. Viovy. 'Mobility of a Reptating Polymer'. Physical Review Letters 69, no. 22 (1992): 3260–63. https://doi.org/10.1103/PhysRevLett.69.3260

[75]Fangman, Walton L. 'Separation of Very Large DNA Molecules by Gel Electrophoresis'. Nucleic Acids Research 5, no. 3 (1978): 653–65. https://doi.org/10.1093/nar/5.3.653

A further innovation, called pulsed-field gel electrophoresis, was developed by Schwartz et al. (1984)[76]. Pulsed-field gel electrophoresis uses an electric field that swaps back and forth between two orientations, each one held for many minutes. DNA moving through a gel orients lengthwise, in parallel with its movement through the gel, digging through the gel like a snake (hence the term "reptation" for this kind of movement). Once the DNA is oriented in the direction of movement, it's moving at some roughly fixed speed regardless of DNA length, for long DNA molecules. But before that happens, shorter and longer DNA molecules behave differently: longer ones take a longer time to orient lengthwise, and until they're oriented they move through the gel more slowly. The idea of pulsed-field gel electrophoresis is to take advantage of that by making the DNA molecules repeatedly reorient[77]:

Figure 1: Schematic diagrams of PFGE instrumentaion.



The electrophoresis device (ED, a) and pulse–oriented electrophoresis (POE, b) instruments are described in detail in Box 1.

(Figure 1 from Herschleb (2007). © publisher)

Orbach et al. (1988) were able to use PFGE to separate chromosomes of length up to $O(10^7)$ bp[78], which is still over an order of magnitude short of most human chromosomes.

I don't understand why there's a barrier around $O(10^7)$ bp, where DNA longer than that can't feasibly be separated using PFGE; this is discussed in e.g. Slater (2009)[79]. There are several other problems with PFGE as a chromosome identification method:

- Existing methods are only able to separate large DNA molecules by using very weak electric fields,

[76]Schwartz, D. C., and C. R. Cantor. 'Separation of Yeast Chromosome-Sized DNAs by Pulsed Field Gradient Gel Electrophoresis'. Cell 37, no. 1 (1984): 67–75. https://doi.org/10.1016/0092-8674(84)90301-5

[77]Herschleb, Jill, Gene Ananiev, and David C. Schwartz. 'Pulsed-Field Gel Electrophoresis'. Nature Protocols 2, no. 3 (2007): 677–84. https://doi.org/10.1038/nprot.2007.94

[78]Orbach, Marc J., Douglas Vollrath, Ronald W. Davis, and Charles Yanofsky. 'An Electrophoretic Karyotype of Neurospora Crassa'. Molecular and Cellular Biology 8, no. 4 (1988): 1469–73. https://doi.org/10.1128/mcb.8.4.1469-1473.1988

[79]Slater, Gary W. 'DNA Gel Electrophoresis: The Reptation Model(s)'. ELECTROPHORESIS 30, no. S1 (2009): S181–87. https://doi.org/10.1002/elps.200900154

and by waiting for many minutes for the DNA to reorient itself after each switch of the electric field direction. This takes many hours or even up to multiple days for a single sorting run. For example, Cox et al. (1990) took 10 days to separate chromosomes of lengths up to $O(10^7)$ bp[80]. That's several times slower than other methods such as staining.

- PFGE as usually practiced always involves stripping proteins from chromatin. (I think the reason is that chromatin proteins make the DNA move slower, and might add variation—i.e. different chromosomes of the same number but with different structural proteins attached might move at significantly different rates, leading to poor separation from adjacent chromosome numbers.) This is commonly done with proteinase K, which even removes histones. Even if this treatment doesn't directly damage DNA, it would disrupt the DNA's epigenomic state, and it might make the DNA unable to function normally in zygogenesis.
- There's likely shear damage to some DNA molecules sorted with PFGE; and the DNA is made more fragile by the protein-stripping treatments. DNA breakage can be partly avoided by tweaking the setup, e.g. by using weaker electric fields. I don't know how low the damage rates can get. As discussed in "Confidence of intact DNA", the rates have to be small—on the order of <1%.
- The sorted chromosomes would have to be extracted from the gel.
- If PFGE is used in bulk, then the disadvantages of bulk methods apply (no index info, so no complementary verification). If PFGE is used on the contents of an individual cell, then the tiny individual DNA molecules would have to be located within a huge blob of gel.

Bakajin et al. (2001) report extremely fast (seconds) sorting of $O(10^5)$ bp DNA using micropillar arrays instead of gel[81]:

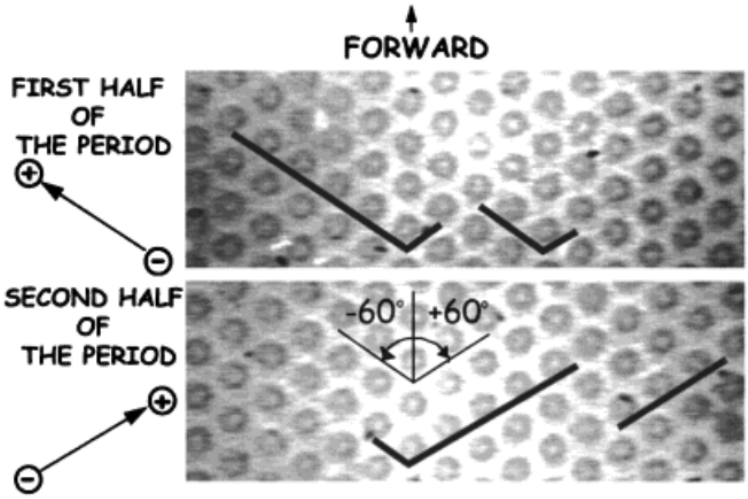


Figure 1. Microscopic view of the hexagonal array. Pillars 2 μm wide were used in these experiments. Cartoon DNA molecules are drawn to illustrate the motion of individual molecules of different lengths. Each period T consists of two pulses aligned along the channels created by the posts in the array, giving a net angle between the two field directions of 120°. Shorter molecules move farther in the array because once they have reoriented along the axis of the field they move in an unhindered straight line for the duration of the pulse. Longer molecules, on the other hand, spend most of the pulse period retracing their paths. In the example shown here, the longer of the two will never advance.

(Figure 1 from Bakajin et al. (2001). © publisher)

But that method doesn't seem to have progressed to larger molecules.

Kim et al. (1995) report very fast (minutes) separation of $O(10^6)$ bp DNA using capillary PFGE[82].

[80]Cox, E C, C D Vocke, S Walter, K Y Gregg, and E S Bain. 'Electrophoretic Karyotype for Dictyostelium Discoideum.' Proceedings of the National Academy of Sciences 87, no. 21 (1990): 8247–51. https://doi.org/10.1073/pnas.87.21.8247

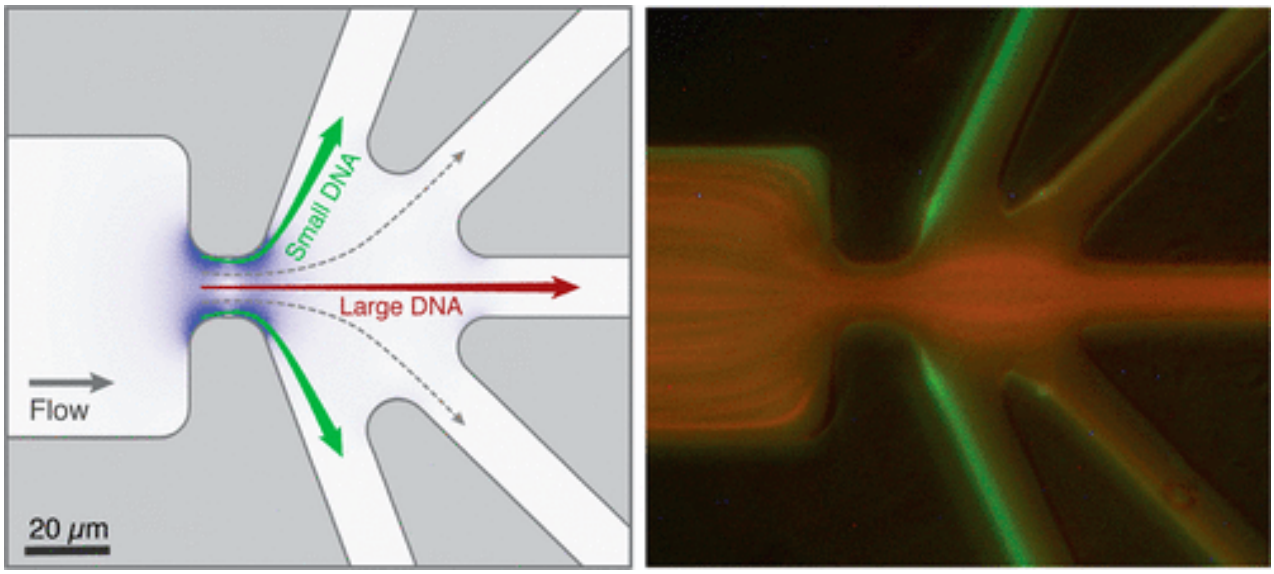
[81]Bakajin, Olgica, Thomas A. J. Duke, Jonas Tegenfeldt, et al. 'Separation of 100-Kilobase DNA Molecules in 10 Seconds'. Analytical Chemistry 73, no. 24 (2001): 6053–56. https://doi.org/10.1021/ac015527o

[82]Kim, Yongseong., and Michael D. Morris. 'Rapid Pulsed Field Capillary Electrophoretic Separation of Megabase Nucleic Acids'. Analytical Chemistry 67, no. 5 (1995): 784–86. https://doi.org/10.1021/ac00101a002

Dielectrophoresis

Dielectrophoresis can also be used to move DNA. It has been used to successfully sort larger objects such as cells; e.g. Yang et al. (2000) were able to sort various types of human leukocytes using dielectrophoresis[83]. The dielectrophoretic force on a particle depends on many aspects of the particle and the applied field (e.g. the frequency of voltage alternation), so in general many different kinds of particles could plausibly be sorted with dielectrophoresis[84][85].

For example, Jones et al. (2017) used dielectrophoresis to sort small and large DNA molecules in fluid buffer[86]:



(Figure from Jones et al. (2017). © publisher)

However, the DNA they sorted was short—only $O(10^3) - O(10^5)$ base pairs, whereas chromosomes are $O(10^8)$ bp. (That's similar performance to Parikesit et al. (2008)[87], but with higher throughput and nicer images.)

Prinz et al. (2002) were able to lyse some Escherichia coli cells and isolate the (circular) chromosome (about 5 million bp) from the lysate[88]. Clausen et al. (2011) were able to use dielectrophoresis to noticeably deflect human chromosomes, but they didn't explore differences between chromosomes and there doesn't seem to be follow-up research[89].

---

[83] Yang, Jun, Ying Huang, Xiao-Bo Wang, Frederick F. Becker, and Peter R. C. Gascoyne. 'Differential Analysis of Human Leukocytes by Dielectrophoretic Field-Flow-Fractionation'. Biophysical Journal 78, no. 5 (2000): 2680–89. https://doi.org/10.1016/S0006-3495(00)76812-3

[84] Pethig, Ronald. 'Review—Where Is Dielectrophoresis (DEP) Going?' Journal of The Electrochemical Society 164, no. 5 (2016): B3049. https://doi.org/10.1149/2.0071705jes

[85] Barasinski, Matthäus, Georg R. Pesch, and Georg Garnweitner. 'Chapter 7 - Electrophoresis and Dielectrophoresis'. In Particle Separation Techniques, edited by Catia Contado. Handbooks in Separation Science. Elsevier, 2022. https://doi.org/10.1016/B978-0-323-85486-3.00009-3

[86] Jones, Paul V., Gabriel L. Salmon, and Alexandra Ros. 'Continuous Separation of DNA Molecules by Size Using Insulator-Based Dielectrophoresis'. Analytical Chemistry 89, no. 3 (2017): 1531–39. https://doi.org/10.1021/acs.analchem.6b03369

[87] Parikesit, Gea O. F., Anton P. Markesteijn, Oana M. Piciu, et al. 'Size-Dependent Trajectories of DNA Macromolecules Due to Insulative Dielectrophoresis in Submicrometer-Deep Fluidic Channels'. Biomicrofluidics 2, no. 2 (2008): 024103. https://doi.org/10.1063/1.2930817

[88] Prinz, Christelle, Jonas O. Tegenfeldt, Robert H. Austin, Edward C. Cox, and James C. Sturm. 'Bacterial Chromosome Extraction and Isolation'. Lab on a Chip 2, no. 4 (2002): 207–12. https://doi.org/10.1039/B208010A

[89] Clausen, Casper Hyttel, Maria Dimaki, Sonia Buckley, and Winnie Edith Svendsen. 'Dielectrophoretic Manipulation of Human Chromosomes in Microfluidic Channels: Extracting Chromosome Dielectric Properties'. BioChip Journal 5, no. 1 (2011): 56–62. https://doi.org/10.1007/s13206-011-5109-0

Hydrodynamic methods

Method	Refined	Price	Time	Identifies numbers	Identifies ho-mologs	Set size	Confident	Saves DNA	Saves epige-netic state	Saves struc-tural proteins
Hydrodyna methods	not much applied to chrs	??	minutes??	hasn't been shown		could be ok using mass and shape?	?	shear forces?	?	?

In general, there are several methods that sort particles by flowing them through fixed structures.

For example, you can have a channel with fluid flowing through it, where the flow isn't uniform—e.g. the flow rate and/or the fluid substance is different at different positions. Such a non-uniform flow will induce shear forces on suspended particles. Those shear forces can somehow make particles of different sizes and shapes separate out from each other:

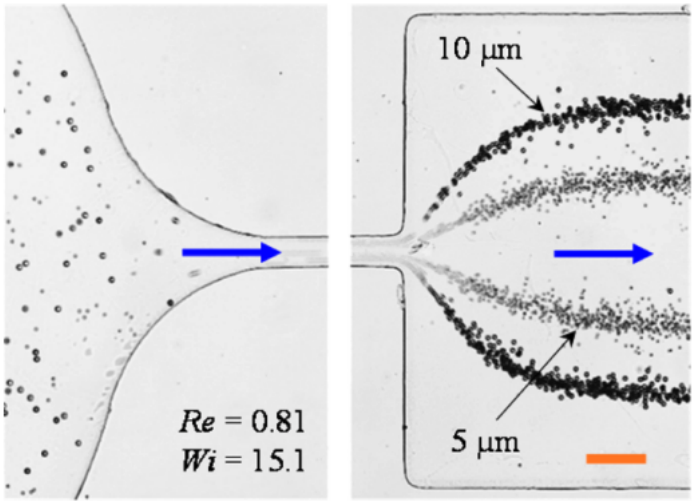


Figure 2. Continuous separation of 5 μm- and 10 μm-diameter spherical particles in 1000 ppm PEO solution through a straight 50 μm-wide and 25 μm-high rectangular microchannel at a flow rate of 250 μL/h: (left) snapshot image at the channel inlet and (right) superimposed image at the outlet expansion. The block arrows indicate the flow direction. The scale bar represents 100 μm.

(Figure 2 from Li et al.[90]. © publisher)

In this vein, Wassberg et al. (2022) ran cell lysates through a series of expansion and contraction steps (wider and narrower channels)[91]:

[90]Li, Di, Xinyu Lu, and Xiangchun Xuan. 'Viscoelastic Separation of Particles by Size in Straight Rectangular Microchannels: A Parametric Study for a Refined Understanding'. Analytical Chemistry 88, no. 24 (2016): 12303–9. https://doi.org/10.1021/acs.analchem.6b03501

[91]Wassberg, Therese R., Mathilde L. Witt, Murat Serhatlioglu, Christian F. Nielsen, Ian D. Hickson, and Anders Kristensen. 'Size-Based Chromosome Separation in a Microfluidic Particle Separation Device Using Viscoelastic Fluids'. EPJ Web of Conferences 266 (2022): 12007. https://doi.org/10.1051/epjconf/202226612007

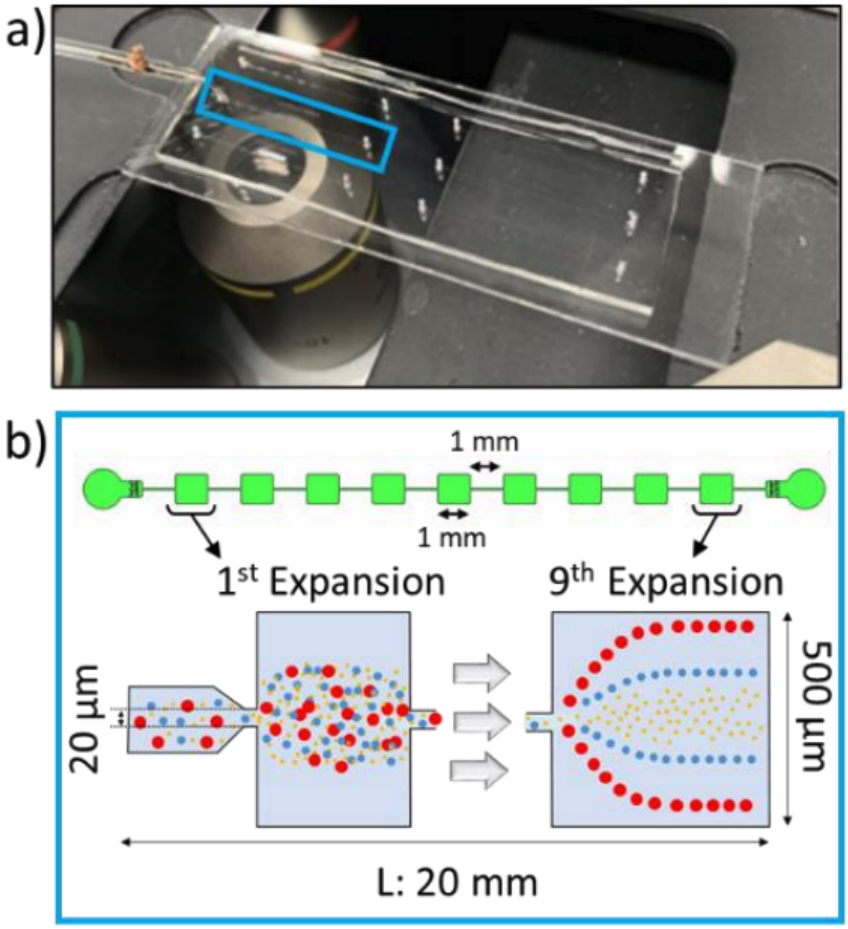


Fig. 1. a) The fabricated PDMS device and experimental setup, and b) schematic illustration of channel design and particle separation.

(Figure 1 from Wassberg et al. © publisher)

They show this image of trajectories of a couple individual chromosomes of different sizes:

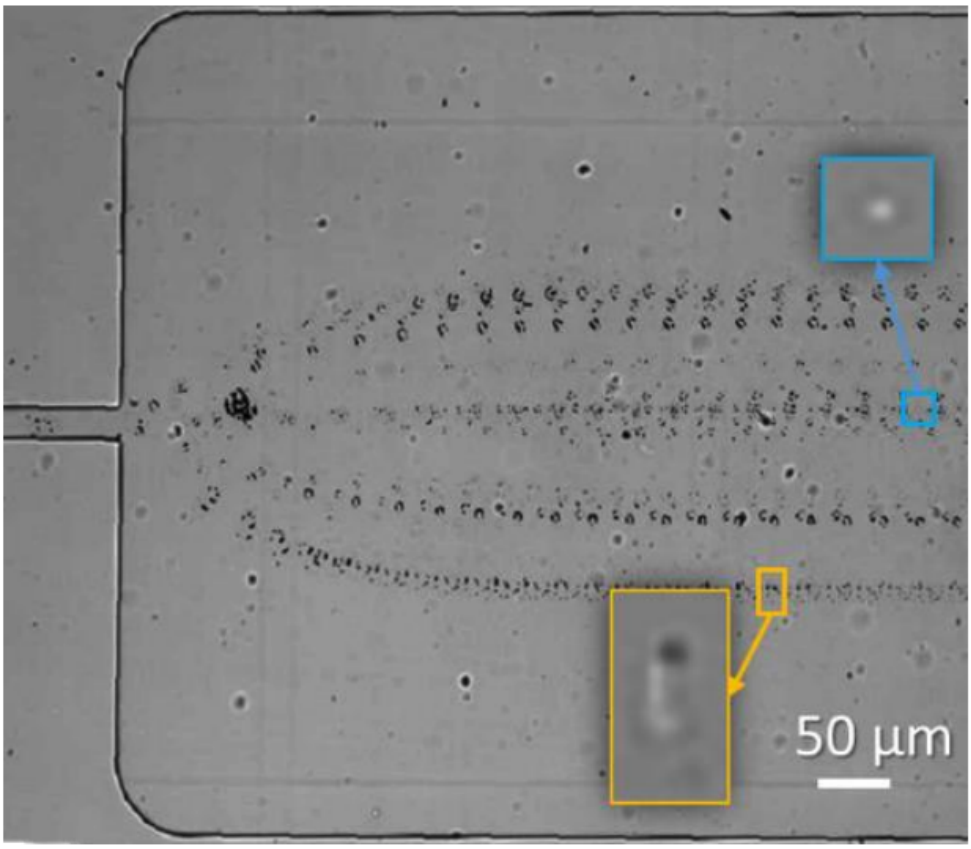


Fig. 3. Image stacked picture of chromosomes, suspended in 9000 ppm PBS/PEO_{600kDa} at 720 µL/h flow rate. The inserted images are the single-frame snapshots (7 times enlarged) from high-speed camera video recordings showing the location of 2.5 µm (blue outline) and 8.1 µm (yellow outline) size chromosomes.

(Figure 3 from Wassberg et al. © publisher)

It's hard to tell how good the separation would be for similarly-sized chromosomes. One imagines that numbers $\{10, 11, 12\}$ would not be separated this way. Even so, this could be a useful way to pre-sort chromosomes as a screening identification method (followed by a confirmation test; see the subsection "Screening and confirmation tests" above).

The same fundamental principle of shear lift was used by Feng et al. (2020) to sort particles by running them through a curving track[92]:

[92]Feng, Haidong, Matthew Hockin, Mario Capecchi, Bruce Gale, and Himanshu Sant. 'Size and Shape Based Chromosome Separation in the Inertial Focusing Device'. Biomicrofluidics 14, no. 6 (2020): 064109. https://doi.org/10.1063/5.0026281

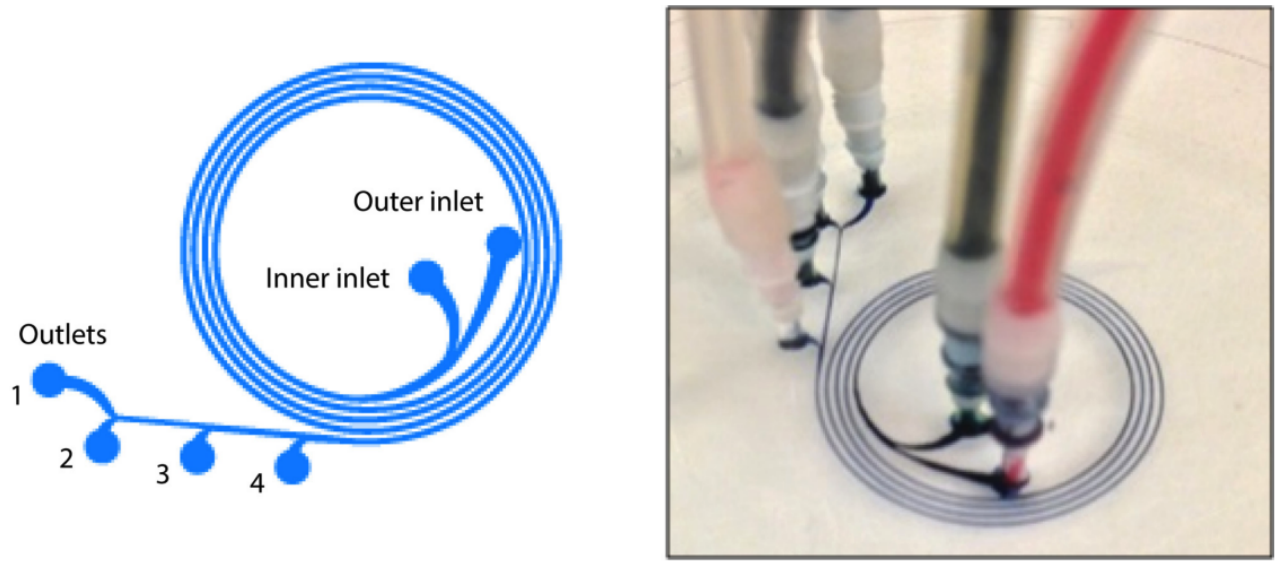
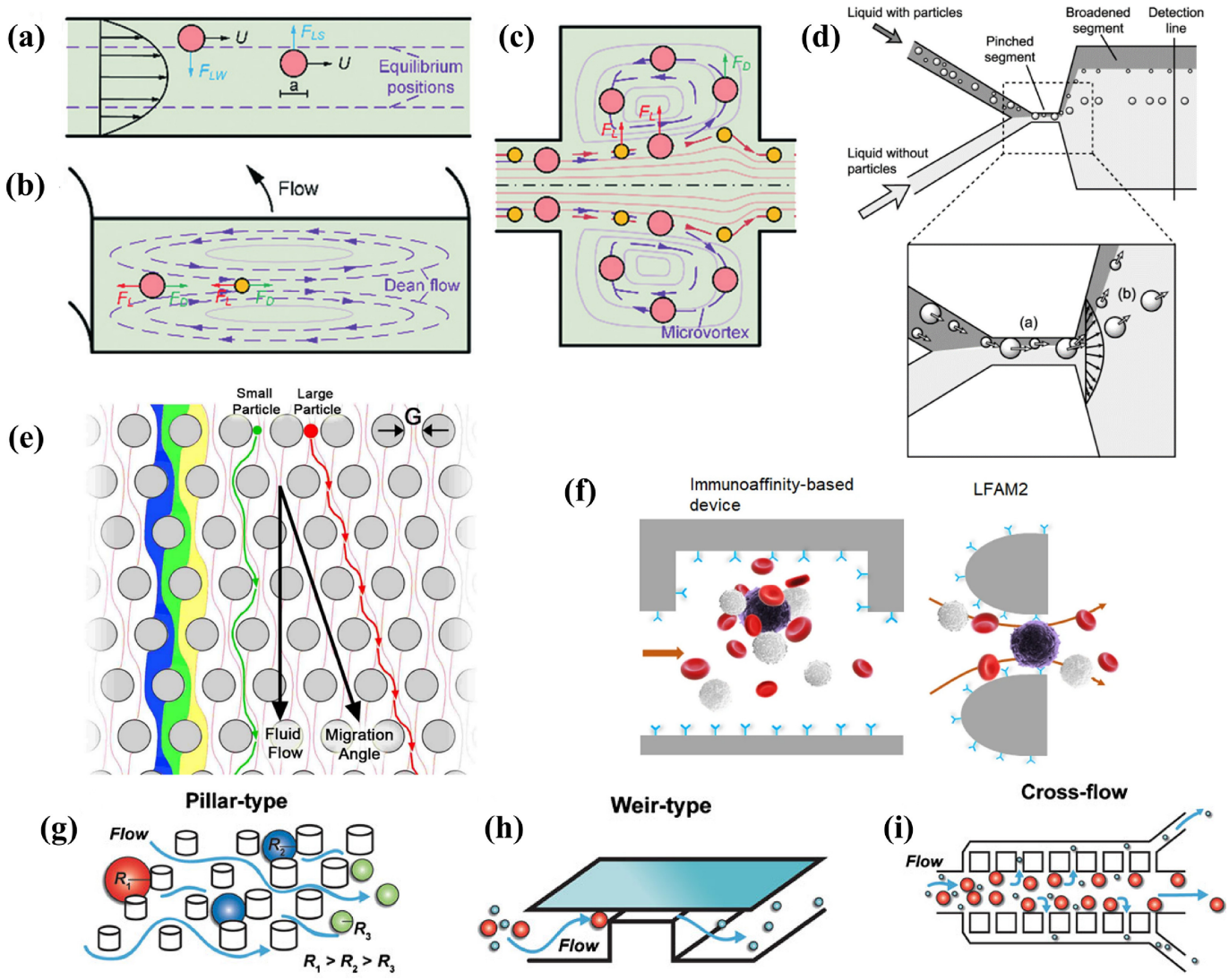


FIG. 2. Schematic diagram and image of the spiral microchannel device.

(Figure 2 from Feng et al. © publisher)

However, their apparatus didn't separate particles very well—seemingly not well enough to separate chromosomes by size.

There are other methods that sort particles by flowing them through fixed structures, such as vortex-forming chambers or pillars that filter particles:



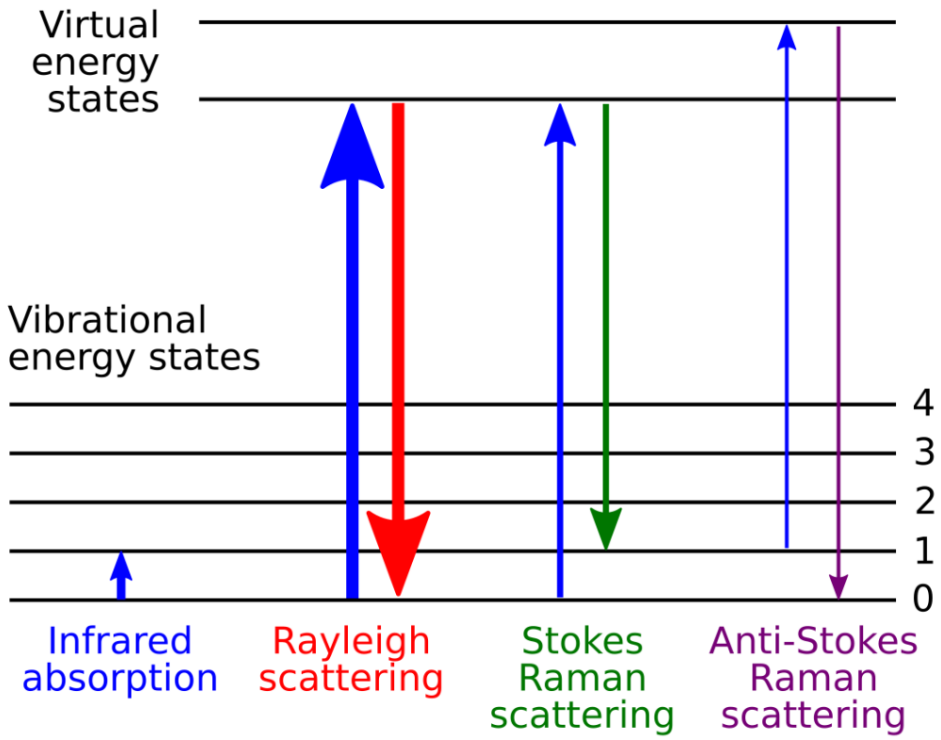
(Figure 3 from Afsaneh and Mohammadi (2022). © publisher)

See Afsaneh and Mohammadi (2022) for more information[93]. I don't know if any of these methods have been applied to chromosomes. All of these methods would be at risk of applying too much shear force to chromosomes, causing them to break.

Raman spectroscopy

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Raman spectrosco	one paper	?	minutes?? but low throughput	claimed		might be precise? claimed 1 vs. 2	might be confident?	laser damage?	?	?

The idea of Raman spectroscopy is that you identify a molecule by how it responds to having a laser shined on it. The molecule absorbs some of the photons of the laser beam, and so the molecule gets unnaturally excited—it stores some energy. Then it unexcites, falling back to a natural state and emitting photons. Sometimes the photons are more energetic or less energetic than the photons that were absorbed, depending which natural unexcited state the molecule falls back into. Exactly how energetic the emitted photons are (i.e. their frequency), will depend on the natural energy states of the molecule. So, different molecules might emit different colors of light, and sometimes you can tell the difference between two molecules based on which photons they emit.



(Visualization by Moxfyre. © publisher)

Researchers have used the Raman scattering principle to identify biological material. E.g. Xie and Li (2003) could distinguish different cells by their Raman spectra[94]:

[93] Afsaneh, Hadi, and Rasool Mohammadi. "Microfluidic Platforms for the Manipulation of Cells and Particles." Talanta Open 5 (August 2022): 100092. https://doi.org/10.1016/j.talo.2022.100092.

[94] Xie, Changan, and Yong-qing Li. 'Confocal Micro-Raman Spectroscopy of Single Biological Cells Using Optical Trapping and Shifted Excitation Difference Techniques'. Journal of Applied Physics 93, no. 5 (2003): 2982–86. https://doi.org/10.1063/1.1542654

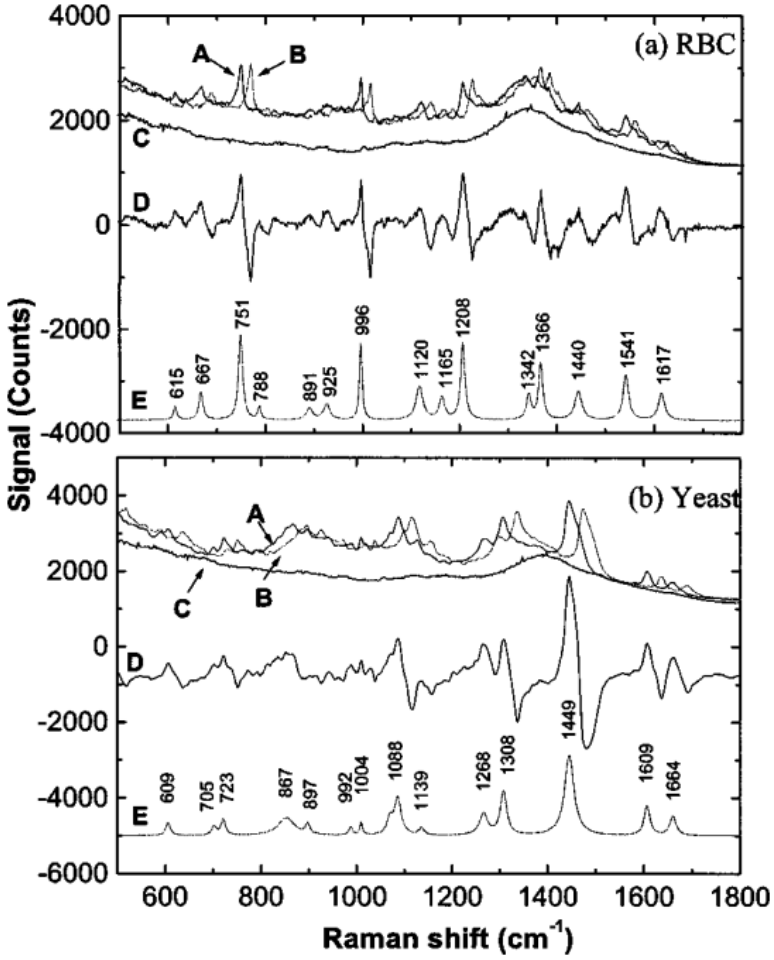


FIG. 4. (a) The different Raman spectrum of a single blood cell trapped in water. Curves A and B are the spectra with an excitation wavelength of 783.0 and 784.2 nm, respectively. (b) The different Raman spectrum of a single living yeast cell trapped in water. Curves A and B are the spectra with an excitation wavelength of 781.3 and 783.1 nm, respectively. The acquisition time is 60 s and the trapping power is 16 mW. Curve C is the background without a cell in the trap. Curve D is the calibrated difference spectrum (A-B). Curve E is the "best guess" reconstructed spectrum.

They used an apparatus like this to shine the laser on the cells and collect the light scattering away from the cells:

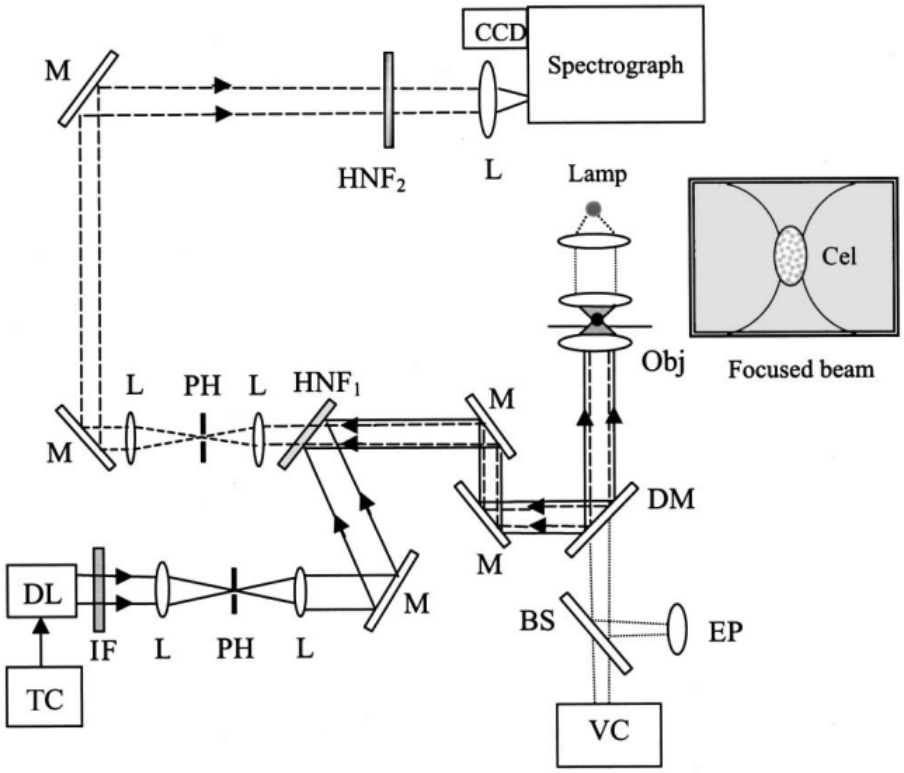


FIG. 1. Experimental setup. A diode laser (DL) beam passes through an interference filter (IF), a pinhole (PH), and is then introduced into an inverted microscope through a dichroic mirror (DM) to form an optical trap. The Raman-scattering light of the trapped particles is collected with the same objective, passes through a holographic notch filter (HNF_1) and a pinhole (PH), and is then focused onto the entrance of an imaging spectrograph equipped with a CCD detector. M—reflection mirror; L—lens; BS—beam splitter; Obj—objective lens; EP—eyepiece; VC—video camera; TC—Temperature controller; and Lamp—Green-filtered xenon illumination light.

Ojeda et al. (2006) reported that they could distinguish Raman spectra of human chromosomes 1, 2, and 3, even though those chromosomes are close to each other in length[95]:

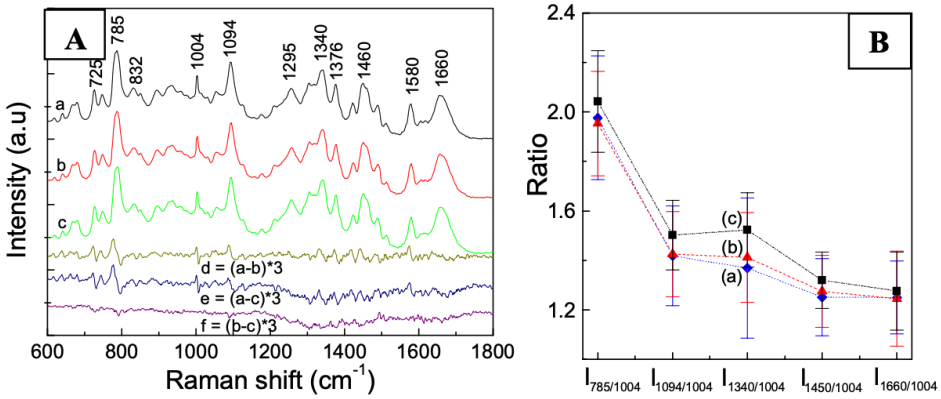


Fig. 3. (a) The averaged Raman spectra of three groups and the differences among them. The different spectra are multiplied by 3 for visualization. a: #1; b: #2, and c: #3. For d,e, and f, each peak is a difference between chromosomes 1 and 2, 1 and 3, and 2 and 3, respectively. (b) The variance among the ratios of Rama peak intensity of main peaks to that at 1004 cm^{-1} of the three chromosomes. The variance was calculated over 22 chromosomes per group.

They appear to get good precision in their classifications:

[95]Ojeda, Jenifer F., Changan Xie, Yong-Qing Li, Fred E. Bertrand, John Wiley, and Thomas J. McConnell. 'Chromosomal Analysis and Identification Based on Optical Tweezers and Raman Spectroscopy'. Optics Express 14, no. 12 (2006): 5385–93. https://doi.org/10.1364/OE.14.005385

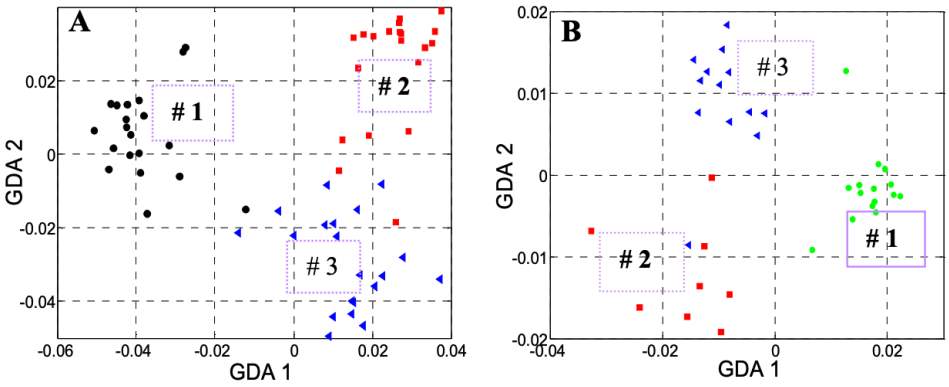


Fig. 4. Generalized discriminate analysis (GDA) plots of all three chromosome numbers. GDA plot using all data collected from all three chromosomes and normalizing the peaks by a chosen standard of 783cm-1. Chromosome 1 is represented as black circles, chromosome 2 as red squares, and chromosome 3 as blue triangles. (a) are the Raman spectra of chromosomes isolated from 6 donors over 12 different days. (b) are Raman spectra from the chromosomes of a single individual over 6 different days.

I don't know whether the differences in spectra are due to different proteins bound to the DNA or due to macro-scale sequence differences (e.g. different counts of G-C pairs vs. A-T pairs). Since they found significant person-to-person differences, bound proteins probably do affect the spectra significantly (as the sequence differences between different people's chromosome 1s are probably negligible).

Raman spectroscopy is somewhat appealing as a chromosome identification method:

- It can plausibly be made non-destructive by using a low-powered laser.
- It can plausibly be accurate enough, just from eyeballing their diagrams.
- The measurement itself (ignoring the surrounding sample preparation) is fast (minutes).

However:

- The setup is somewhat complex. It requires a complicated apparatus, and requires novel calibration; we'd have to collect a bunch of staining data to learn to discriminate between chromosomes based on their Raman spectra.
- I don't know if the method actually works. The 2006 paper has about 50 citations on Google Scholar, but none of them seem to be replicating / extending / applying their work to identify chromosomes. (In 2007 the authors reaffirmed their conclusions in a response to a reviewer[96].) I couldn't contact any of the authors.
- I don't know if the lasers cause DNA damage.
- I think they applied the laser measurements before fixing chromosomes in an acidic buffer; but I'm not exactly sure of the meaning of how they treated the chromosomes before measuring.

(Note that this method is a combined method in two respects: They sort chromosomes by size and centromere location, in order to focus on chromosomes 1, 2, and 3. Also, they learn to discriminate between chromosomes 1, 2, and 3 from their Raman spectra by comparing to staining (G-banding).)

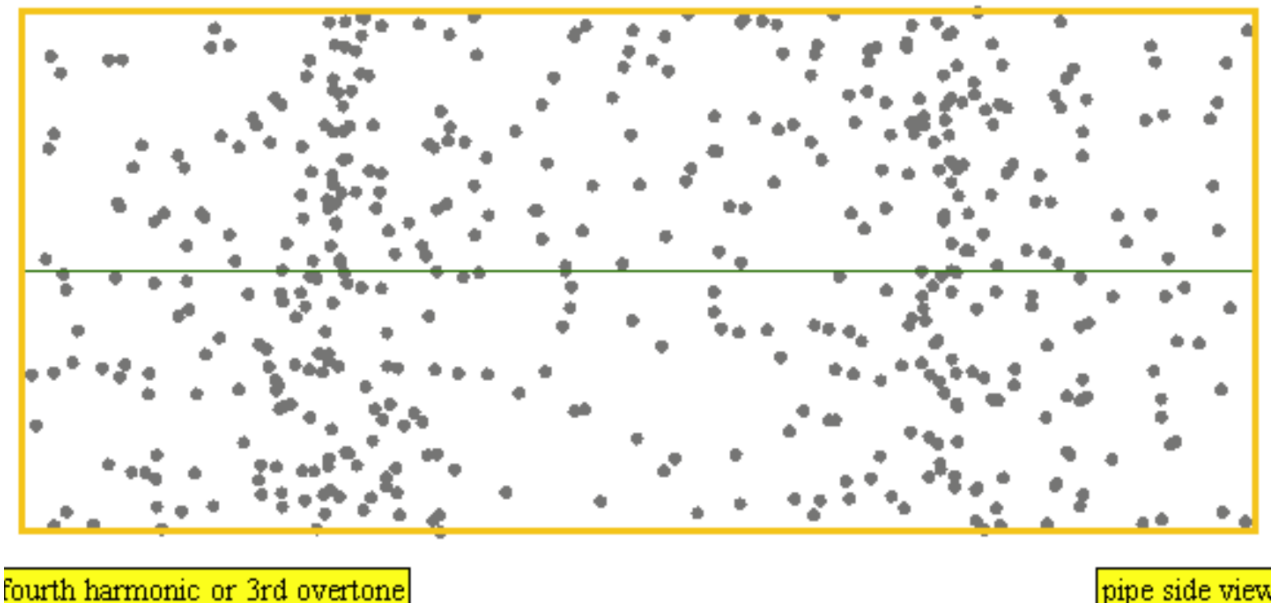
Acoustic methods

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Scanning acoustic microscopy	one paper	?	minutes?? but low throughput			might be about as precise as staining?	might be confident?	?	mild fixation, maybe not needed	mild fixation, maybe not needed
Acoustic sorting	not applied to chrs	??	minutes or seconds??	hasn't been shown		?	might be confident?	shear?		

[96] Ojeda, Jenifer F., Changan Xie, Yong-Qing Li, Fred E. Bertrand, John Wiley, and Thomas J. McConnell. 'Chromosomal Analysis and Identification Based on Optical Tweezers and Raman Spectroscopy: Reply'. Optics Express 15, no. 10 (2007): 6000–6002. https://doi.org/10.1364/OE.15.006000

Acoustic force spectroscopy

Sound waves in a liquid somehow push on particles in the liquid. If the acoustic waves form a standing wave (acoustic resonance), there are nodes where acoustic pressures will be symmetric:



(Gif by Lookang (acknowledging Juan M. Aguirregabiria, Francisco Esquembre), source).

In acoustic force spectroscopy, a particle is somehow located in an acoustic pressure node, and it is tethered to a sensor. The sensor measures how much the particle is being pulled around. How much the particle is being pulled around, and how that depends on the acoustic wave, and how that changes over time (e.g. by breaking the particle), says something about the particle. An illustration:

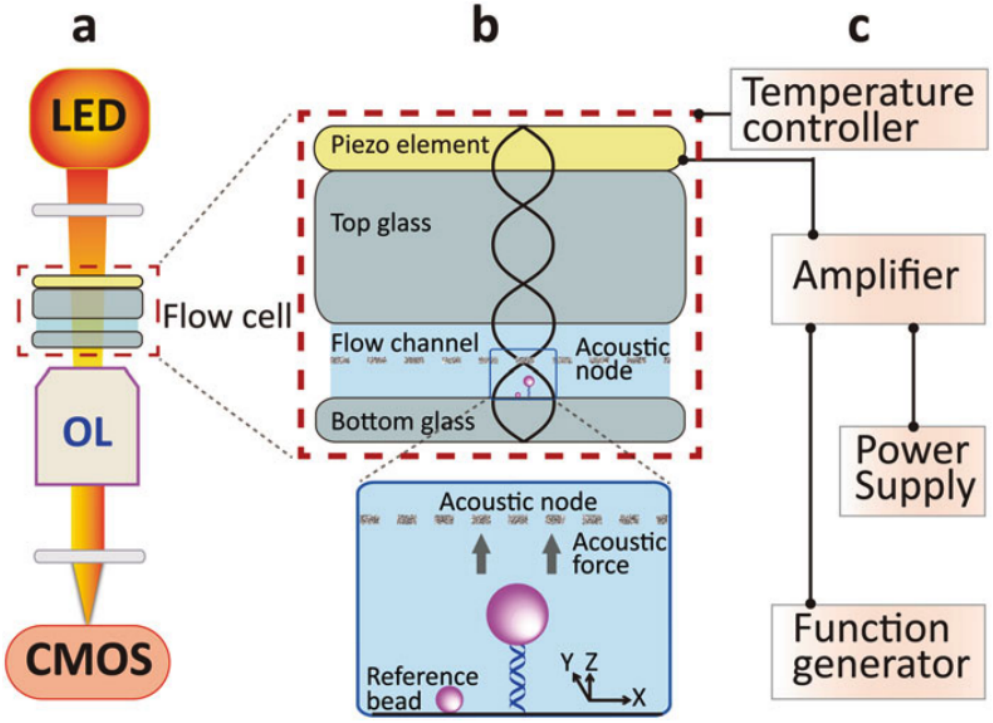


Fig. 1 Illustration of the principle of Acoustic Force Spectroscopy. (**a**) The acoustic force spectroscopy instrument consists of a flow cell, an inverted microscope visualizing targets with objective lens (OL), temperature controller connected to the AFS-Chip holder, a CMOS digital camera and 660 nm LED light source. (**b**) AFS flow cell consists of the piezo element and two glass slides with a fluidic channel in between. The acoustic wave generated by the piezo travels through the top glass to the bottom glass and the bottom glass as a reflector reflects the acoustic wave, producing the standing wave over the flow cell. The acoustic standing waves carry pressure profile, which generate acoustic forces. The tethered particles in the flow channel exposed to the acoustic force are driven in the direction of the acoustic pressure node. (**c**) The temperature controller independently connects to the AFS-Chip metal holder. Both power supply and function generator connect to the amplifier, which controls the piezo element

(Figure 1 from Lin et al. (2018)[97]. © publisher)

In theory you could tell which chromosome you have by how it responds to acoustic forces like that. The forces (e.g. 10s of piconewtons) are in the same ballpark (perhaps one order of magnitude greater) as the forces experienced by metaphase chromosomes during anaphase (wherein the spindle fibers pull on kinetochores attached to chromosomes's centromeres to segregate the two genome copies into daughter cells)[98]. So, the chromosomes might not tend to break. Acoustic force spectroscopy has been used to somehow test DNA molecules ($O(10^4)$ bp)[99]. But, I'm not aware of acoustic force spectroscopy being used to distinguish chromosomes.

Scanning acoustic microscopy

Sound waves bounce off of objects, and do so with greater or lesser intensity and dispersion, depending on the object. An acoustic microscope shoots an acoustic wavefront (a pressure wave) that converges to a point on a sample, and then measures the reflected sound wave:

[97]Lin, Szu-Ning, Liang Qin, Gijs J. L. Wuite, and Remus T. Dame. 'Unraveling the Biophysical Properties of Chromatin Proteins and DNA Using Acoustic Force Spectroscopy'. In Bacterial Chromatin: Methods and Protocols, edited by Remus T. Dame. Springer, 2018. https://doi.org/10.1007/978-1-4939-8675-0_16

[98]Keizer, Veer I. P., Simon Grosse-Holz, Maxime Woringer, et al. 'Live-Cell Micromanipulation of a Genomic Locus Reveals Interphase Chromatin Mechanics'. Science 377, no. 6605 (2022): 489–95. https://doi.org/10.1126/science.abi9810.

[99]Sitters, Gerrit, Douwe Kamsma, Gregor Thalhammer, Monika Ritsch-Marte, Erwin J. G. Peterman, and Gijs J. L. Wuite. 'Acoustic Force Spectroscopy'. Nature Methods 12, no. 1 (2015): 47–50. https://doi.org/10.1038/nmeth.3183

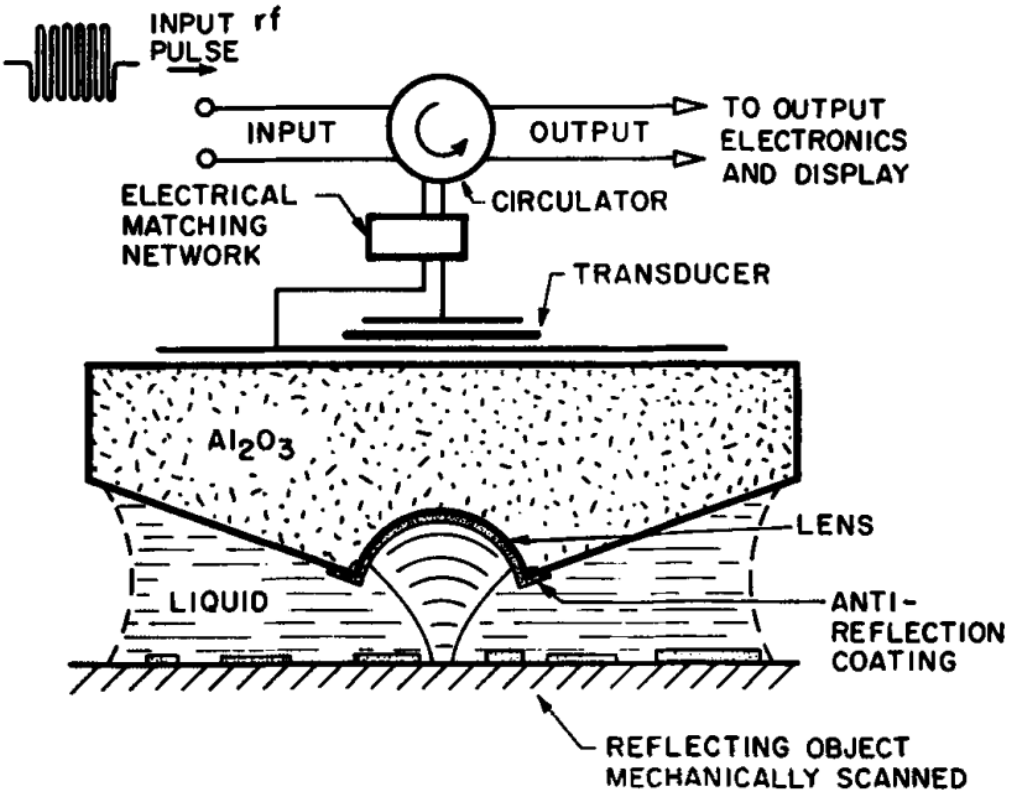


Fig. 1. Sketch of components for reflection acoustic microscopy. Object is mechanically scanned through a focused acoustic beam. The magnitude of the detected acoustic reflection determines the brightness of a picture element in the acoustic micrograph.

(Figure 1 from Rugar et al. (1980). © publisher)

Rugar et al. (1980) applied an acoustic microscope to image human chromosomes, scanning across the sample[100]. They imaged chromosomes treated with trypsin and Giemsa stain:

[100]Rugar, D., J. Heiserman, S. Minden, and C. F. Quate. 'Acoustic Microscopy of Human Metaphase Chromosomes'. Journal of Microscopy 120, no. 2 (1980): 193–99. https://doi.org/10.1111/j.1365-2818.1980.tb04135.x

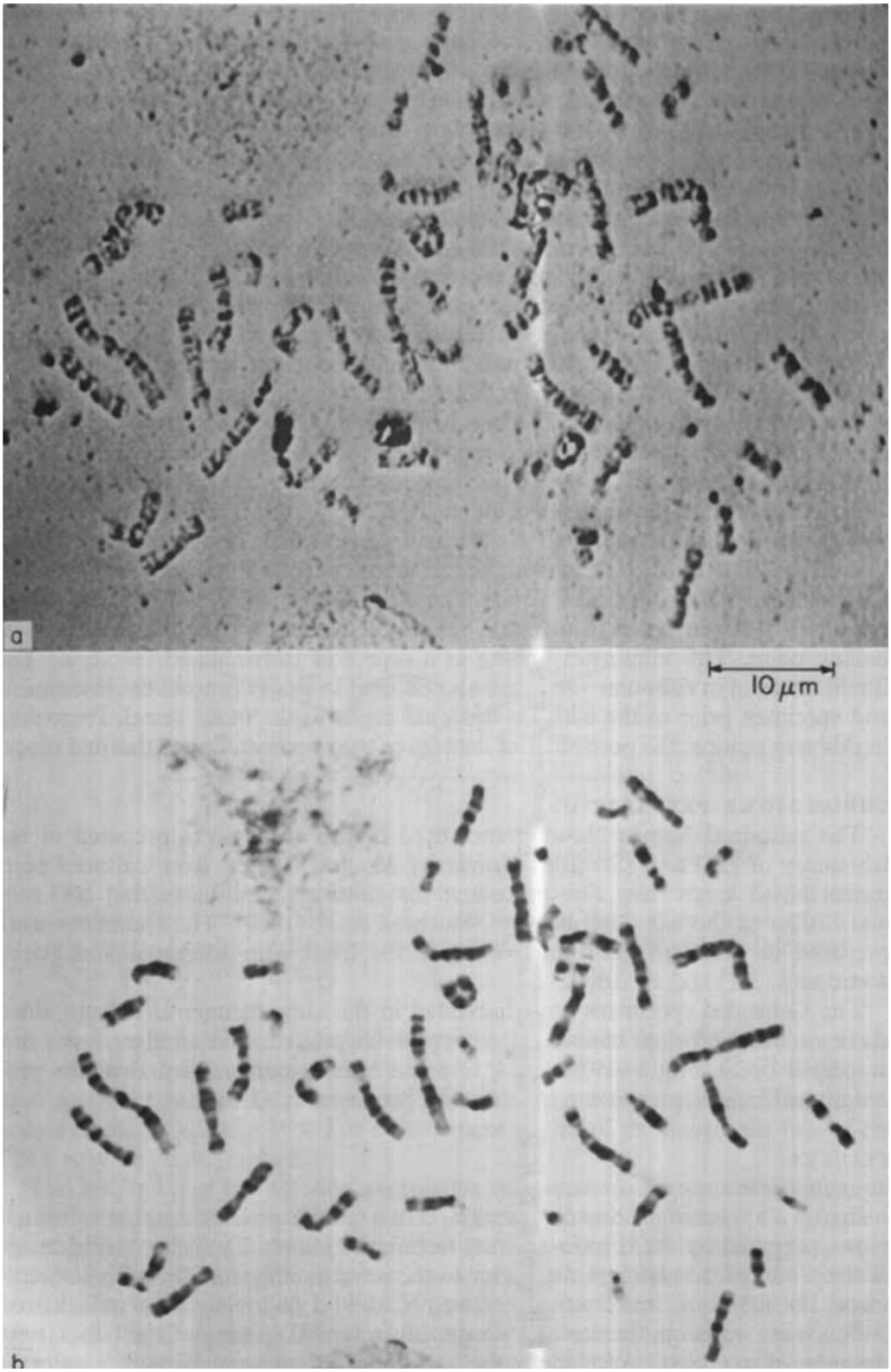


Fig. 2. (a) Acoustic and (b) optical images of a G-banded spread of metaphase human chromosomes prepared by a trypsin–Giemsa procedure. The acoustic image shows distinct G-bands similar to those seen optically. (× 1600.)

(Figure 2 from Rugar et al. (1980). © publisher)

The top image shows the acoustic image of the stained chromosomes. Remarkably, the G-banding pattern is still visible. It's not the same, and it seems lower quality, but it's visible. This could be because the G-bands are formed by Giemsa stain binding to regions where there are fewer proteins (roughly, the heterochromatin regions), either because trypsin removed more proteins from the DNA, or because there were fewer proteins there in the first place. (Indeed, histones and other proteins would be less dense in heterochromatin regions.) Rugar et al. also imaged *unstained* chromosomes with the scanning acoustic microscope:

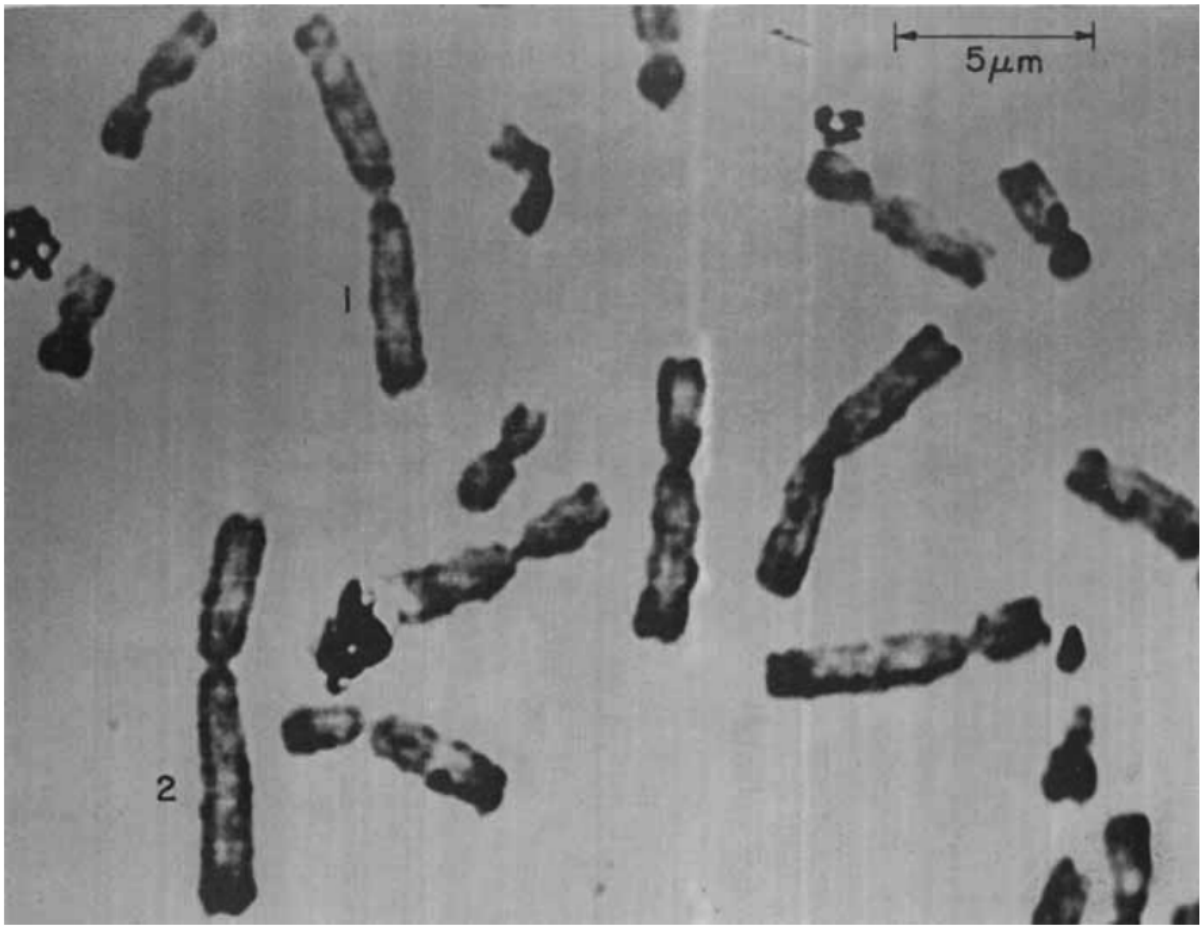


Fig. 4. Acoustic image of unstained chromosomes. The acoustic markings are less distinct than markings in stained specimens and no longer appear as well defined bands. Chromosomes 1 and 2 are identified. (× 4000.)

(Figure 4 from Rugar et al. (1980). © publisher)

The light-dark pattern is fuzzier than the G-bands, and isn't the same pattern. It's possible that this corresponds to the presence of a greater or smaller amount of protein bound to the DNA in different regions, as in the image of the stained chromosomes. If that's the case, then we'd expect those discolorations to be characteristic of chromosome number, since the pattern of euchromatin and heterochromatin regions is characteristic of chromosome number. So, this would potentially be a chromosome numbering method. However, Rugar et al. state "The acoustic markings vary among different chromosomal pairs and even between homologous chromosomal pairs.", which seems to negate that possibility.

I don't know whether an acoustic microscope like this would damage chromatin. Rugar et al. use liquid argon at 85 Kelvin to increase the acoustic resolution, which might be inconvenient. (It might be possible to get away with using water—I think the resolution would only be cut in half.)

There has been very little followup work on this method, so it might be unworkable, and in any case is not well-developed. Some sources state that acoustic microscopy can match or even exceed the resolution of optical microscopy[101][102].

Note that these chromosomes are still fixed with a methanol and acetic acid wash, so their proteins were probably somewhat loosened, though I don't know how much.

[101] Maev, Roman Gr. Acoustic Microscopy: Fundamentals and Applications. John Wiley & Sons, 2008

[102] Strohm, Eric M., Michael J. Moore, and Michael C. Kolios. 'High Resolution Ultrasound and Photoacoustic Imaging of Single Cells'. Photoacoustics 4, no. 1 (2016): 36–42. https://doi.org/10.1016/j.pacs.2016.01.001

Acoustic sorting

Using standing acoustic waves, Laurell et al. (2006) were able to sort 3 µm plastic beads from 8 µm ones[103]:

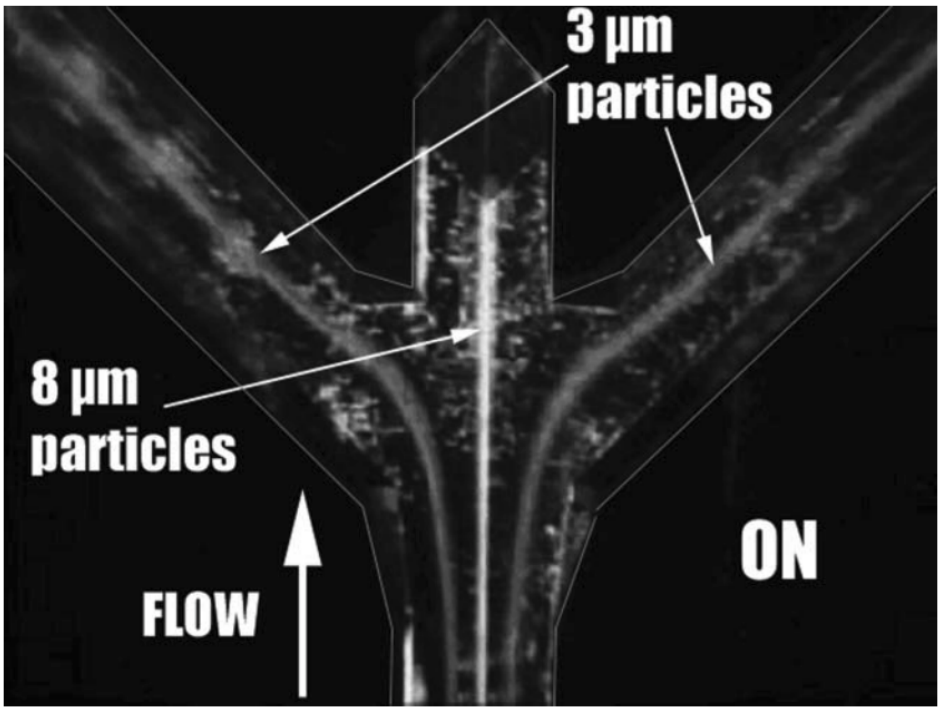


Fig. 31 Microscope image of the situation when the ultrasound is turned on. The 8 µm particles are focused in the fundamental resonance pressure node and exit through the centre outlet while the 3 µm particles are gathered in the first harmonic pressure nodes and exit through the side outlets.

(Figure 31 from Laurell et al. © publisher)

I'm not sure how strong the forces involved are, but Laurell et al. seem to suggest in the piconewton range, which would be acceptable.

A surface acoustic wave apparatus is illustrated in Collins et al. (2016)[104]:

[103]Laurell, Thomas, Filip Petersson, and Andreas Nilsson. 'Chip Integrated Strategies for Acoustic Separation and Manipulation of Cells and Particles'. Chemical Society Reviews 36, no. 3 (2007): 492–506. https://doi.org/10.1039/B601326K.

[104]Collins, David J., Adrian Neild, and Ye Ai. 'Highly Focused High-Frequency Travelling Surface Acoustic Waves (SAW) for Rapid Single-Particle Sorting'. Lab on a Chip 16, no. 3 (2016): 471–79. https://doi.org/10.1039/C5LC01335F.

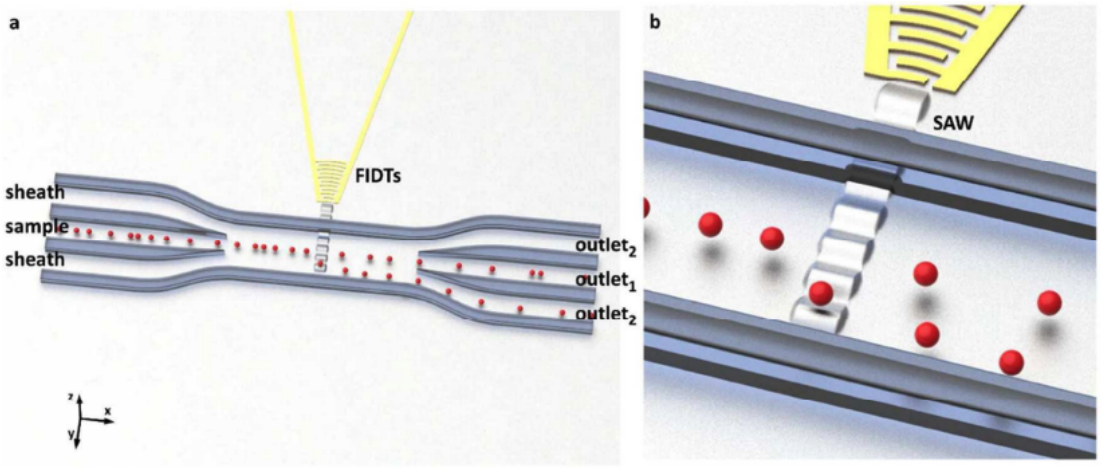
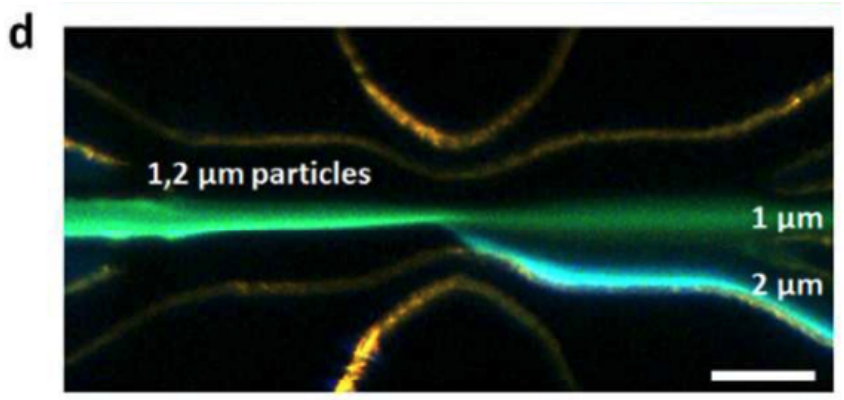


Figure 1: Schematic illustration of the device design. (a) Sheath flow is used to hydrodynamically focus the sample prior to the focused interdigital transducers (FIDTs), which generate the focused surface acoustic wave (SAW) that translates selected particles. (b) Highly focused SAW is used to sort at the single-particle level, where the width of the focused SAW is on the order of 10's of µm.

(Figure 1 from Collins et al. © publisher)

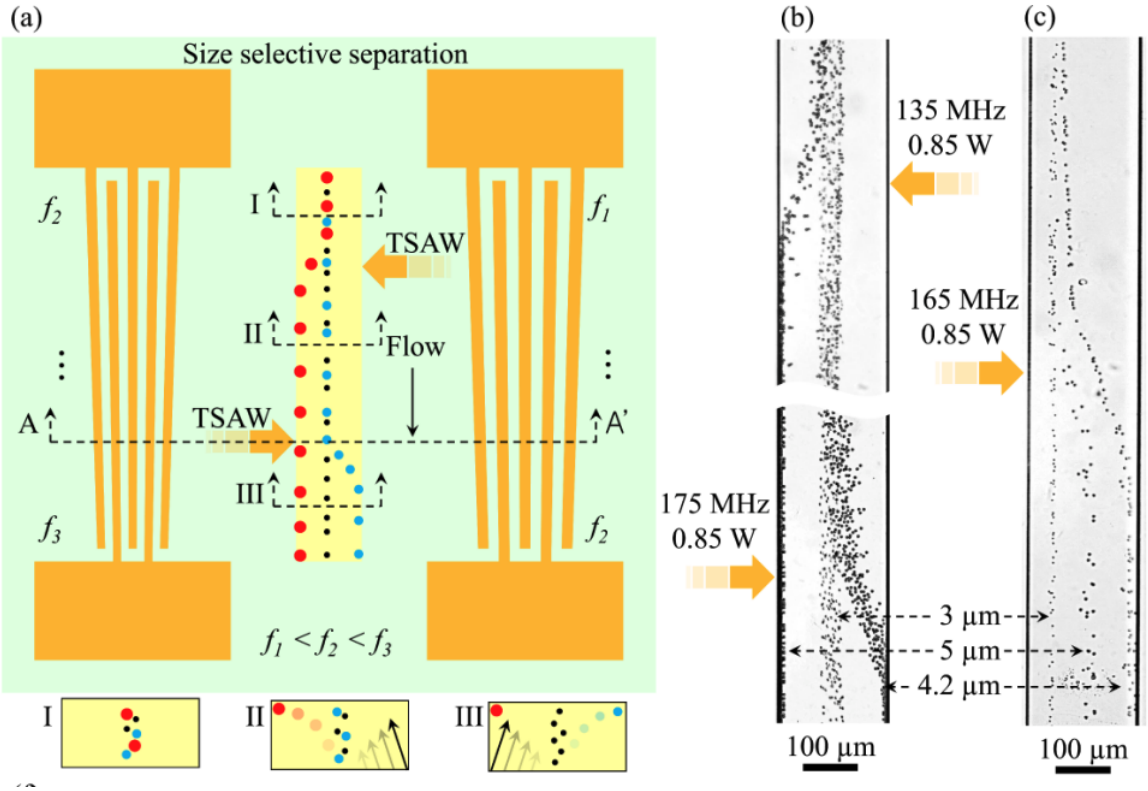
They were able to sort 2 µm particles from 1 µm particles:



(Figure 5 from Collins et al. © publisher)

Another 2015 paper by Destgeer et al. was able to separate particles using a variant of the surface acoustic wave setup[105]:

[105]Destgeer, Ghulam, Byung Hang Ha, Jinsoo Park, Jin Ho Jung, Anas Alazzam, and Hyung Jin Sung. 'Microchannel Anechoic Corner for Size-Selective Separation and Medium Exchange via Traveling Surface Acoustic Waves'. Analytical Chemistry 87, no. 9 (2015): 4627–32. https://doi.org/10.1021/acs.analchem.5b00525.



(Figure 1 from Destgeer et al. © publisher)

They were able to separate polystyrene particles with diameters of 3 µm, 4.2 µm, and 5 m, and with a distinct device, they could separate particles with diameters of 3 µm, 5 µm, and 7 m. The best ratio is $5/4.2 \approx 1.2$, i.e. a 20% difference; that's not yet enough to distinguish chromosomes well enough.

In general, it seems that surface acoustic wave sorting methods can get high purity sorting (>95%), but so far only with large size gaps (>20%)[106]. Another paper was able to sort particles with the same diameter purely using density differences, but only with purities <95%[107]. I did not find examples of chromosomes being sorted this way.

Some more identification methods

I have only cursorily investigated these methods.

Electron microscopy

In general, these methods involve passing electrons through a sample, rather than photons. There are several variants (scanning transmission EM, cryogenic EM, etc.). I don't know what preparation is needed for these methods, what damage would be involved, what they can and cannot resolve, and so on. Here's an example image:

[106]Fan, Yanping, Xuan Wang, Jiaqi Ren, Francis Lin, and Jiandong Wu. 'Recent Advances in Acoustofluidic Separation Technology in Biology'. Microsystems & Nanoengineering 8, no. 1 (2022): 94. https://doi.org/10.1038/s41378-022-00435-6.

[107]Liu, Guojun, Wanghao Shen, Yan Li, et al. 'Continuous Separation of Particles with Different Densities Based on Standing Surface Acoustic Waves'. Sensors and Actuators A: Physical 341 (July 2022): 113589. https://doi.org/10.1016/j.sna.2022.113589.

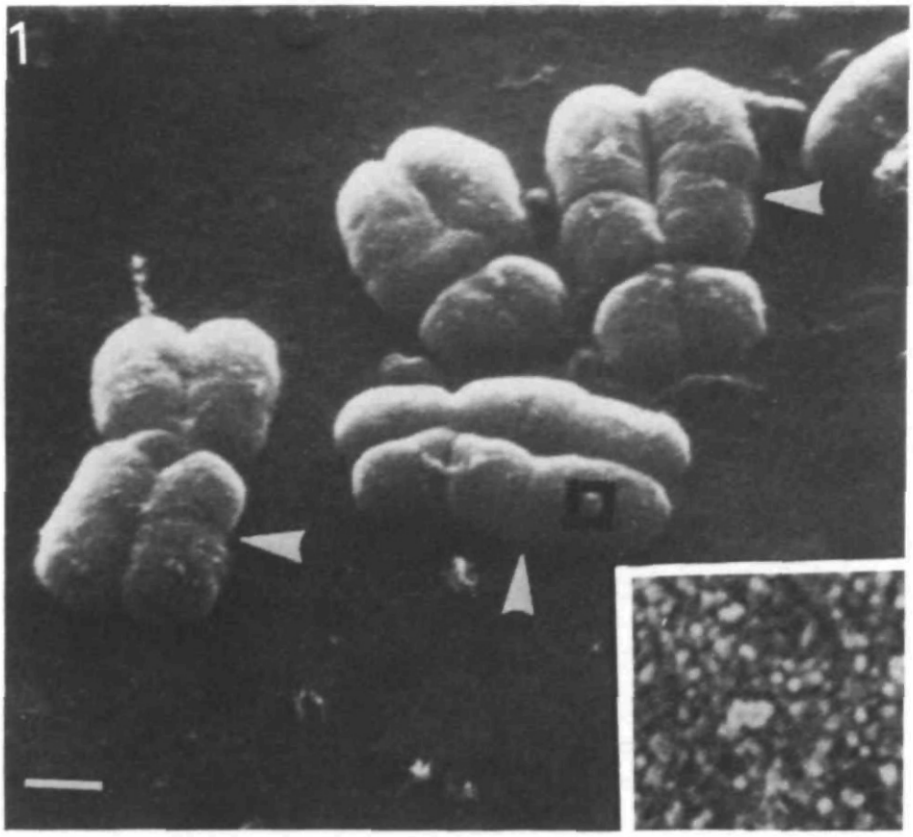


Fig. 1. Non-banded, uncoated human metaphase chromosomes, processed for SEM by the osmium-TCH technique, after initial preparation for LM cytogenetic analysis The chromosomes show a three-dimensional cylindrical profile with well-defined chromatids and centromeres. A smooth surface is shown with occasional indentations along the chromatids (arrows). At higher magnifications (inset) the surface morphology appears irregular. Bar, 1 μm. ×7000. Inset, ×50000.

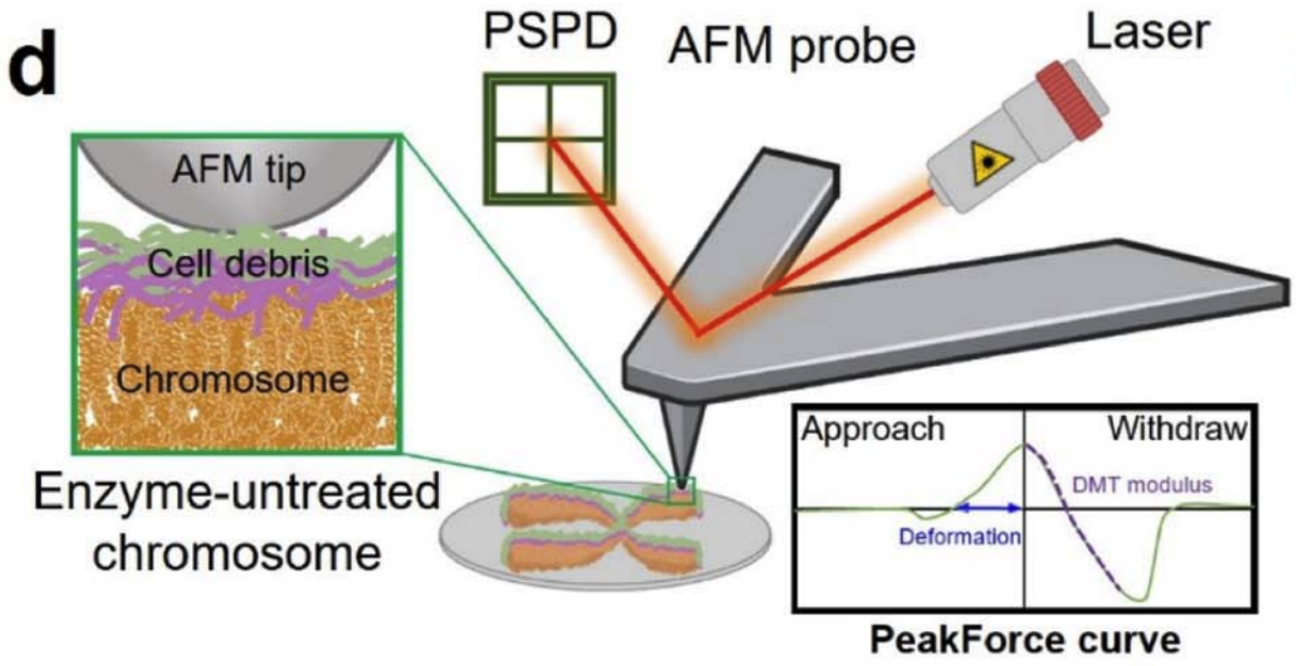
(Figure 1 from Harrison et al.(1982)[108]. © publisher)

It seems like one should be able to at least identify chromosome number from such an image. But imaging a chromosome this way is presumably costly.

Atomic force microscopy

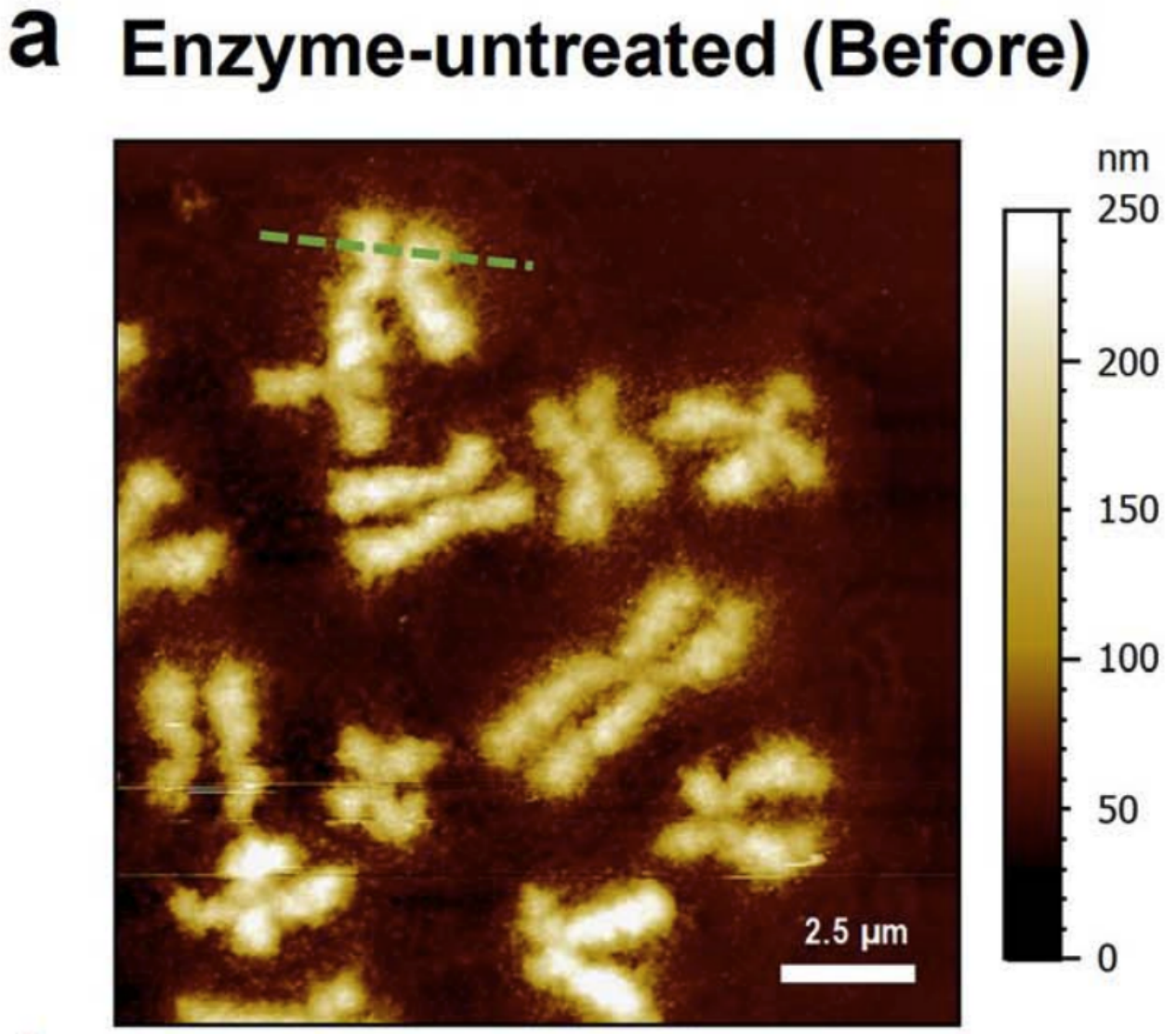
Atomic force microscopy works by physically touching a sample with a metal tip, and measuring how much the metal tip is pushed up. This way, you detect bumps in the sample:

[108]Harrison, Christine J., Terence D. Allen, Martin Britch, and Rodney Harris. 'High-Resolution Scanning Electron Microscopy of Human Metaphase Chromosomes'. Journal of Cell Science 56, no. 1 (1982): 409–22. https://doi.org/10.1242/jcs.56.1.409.



(Figure 3d from Roh et al. (2023)[109]. © publisher)

Roh et al. measured undigested chromosomes using an atomic force microscope, obtaining images like this:



[109]Roh, Seokbeom, Taeha Lee, Da Yeon Cheong, Yeonjin Kim, Soohwan Oh, and Gyudo Lee. 'Direct Observation of Surface Charge and Stiffness of Human Metaphase Chromosomes'. Nanoscale Advances 5, no. 2 (2023): 368–77. https://doi.org/10.1039/D2NA00620K.

(Figure 2a from Roh et al. (2023). © publisher)

This specific image doesn't seem all that much better than light microscopy, but in theory AFM isn't constrained by limits to resolution due to optical diffraction. I don't know other parameters of AFM (damage, speed, etc.); presumably it is too costly.

Detecting allelic proteins or RNAs

In theory it might be possible to do setwise homolog identification by interacting with the proteins or RNAs present in a cell.

See "Sequencing post-meiotic RNA" for more on this idea. That subsection is in the section on indirect identification methods, because it discusses dissociating the cytoplasm (containing proteins and RNAs) from the chromatin. But, the dissociation step might be an inessential step. The version that does direct identification would work like this:

1. Label some proteins/RNAs
   - For a direct method, you do this in vivo, i.e. in the live cell without destroying it.
   - You could label proteins/RNAs that are on the surface of the cell, and/or that are in the inside of the cell (e.g. in the cytoplasm or other compartments).
   - You label using any sort of label that:
     - can be targeted to differences in proteins or RNAs that are downstream of allelic differences, and
     - can be sensed or manipulated from outside of the cell, e.g. fluorophores or magnetic beads.
2. Sort the cells by whether they have the labels you're looking for. (This corresponds to having haploblocks with labeled markers that stake out desired segments of the two paternal chromosomes.)

(This would be a direct method because you didn't fully separately operate on the detected items and the target chromosomes.)

I don't know if this would be practical. Proteins might only weakly indicate genotype; see "Sequencing post-meiotic RNA". RNAs could work, and this might be easier than the indirect method, because at least you skip the lysing step, and you can just directly sort the cell. On the other hand, it's harder—maybe much harder—to make many dozens of allele-specific RNA binders and deliver them inside the cell.

There may not be enough allele-specific items on the surface of sperm to do much homolog identification. Surface proteins have been used to sort sperm based on X/Y content[110], or (I think) based on presence of a mutation-associated protein (which would not be the case for most allele variations)[111].

Even more direct methods

I haven't looked into these.

- *Magnetophoresis.*
  - Similar to electrokinesis, these methods use magnetic fields. See e.g. Munaz et al. (2018)[112].
- *Super-resolution microscopy.*
  - These are methods for getting around the diffraction limit on resolution for light imaging.
  - E.g. Minflux.
- *X-rays.*
  - E.g. nanotomography.
  - E.g. Bhartiya et al. (2021) report getting a karyotype by X-raying human chromosomes[113].
- *Magnetic force microscopy.*
- *Diamond quantum sensors.*
  - See Wu et al. (2016)[114].

[110]Khirbat, Richa, Trilok Nanda, Aman Kumar, et al. 'H-Y Antigen-Based Immuno-Segregation Strategies for Bovine X Sperm Enrichment'. Theriogenology 243 (September 2025): 117474. https://doi.org/10.1016/j.theriogenology.2025.117474.

[111]Adenmosun, Olumide O. 'Genotypic Sperm Sorting: A Less Invasive "ART" to Prevent Genetic Disorders in Newborns'. Ph.D., Florida Atlantic University, 2021. https://www.proquest.com/docview/2572595676/abstract/66584465C63543E3PQ/1.

[112]Munaz, Ahmed, Muhammad J. A. Shiddiky, and Nam-Trung Nguyen. "Recent Advances and Current Challenges in Magnetophoresis Based Micro Magnetofluidics." Biomicrofluidics 12, no. 3 (2018): 031501. https://doi.org/10.1063/1.5035388.

[113]Bhartiya, Archana, Darren Batey, Silvia Cipiccia, et al. "X-Ray Ptychography Imaging of Human Chromosomes After Low-Dose Irradiation." Chromosome Research 29, no. 1 (2021): 107–26. https://doi.org/10.1007/s10577-021-09660-7.

[114]Wu, Yuzhou, Fedor Jelezko, Martin B Plenio, and Tanja Weil. "Diamond Quantum Devices in Biology." Angewandte Chemie International Edition 55, no. 23 (2016): 6586–98. https://doi.org/10.1002/anie.201506556.

Indirect chromosome identification methods

This section gives more information about indirect chromosome identification methods, i.e. ones that don't directly interact with the specific chromosomes that are identified by the method. (See the subsections "Direct and indirect chromosome identification" and "Indirect chromosome identification".)

Reprinting the summary table (see the subsection "Criteria for identification methods" for the meaning of the criteria in this table):

Indirect methods										
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Diploid culture		buffer for ~1 week	days or more	(does tell euploidy)		diploid		mostly fine but accumulates mutation?	disrupts	
Haploid culture	a few papers; logically should work	? needs eggs; low-quality ok	days or more	(does tell euploidy)		haploid; combines well with number ID		mostly fine but accumulates a bit of mutation?	disrupts; unsure how much	
Sequencing post-meiot RNA	1 paper; logically should work	? hard to develop	a day or so?	(does tell euploidy)		haploid; combines well with number ID				
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Chromoso complemer identificati	not shown; logically should work	? hard to develop	minutes?		combines with set-wise homolog ID	1				
Meiotic complemer identificati	1 paper; logically should work	medium unit price??	days?	(does tell euploidy)		haploid; combines well with number ID	much less confident with < 3 meiotic cousins		(creates it; or else disrupts)	(creates them)
Targeted complemer elimination	?	?	hours?		with more design work?	1		probably ok?	? may require fixation like for CRISPR	? may require fixation like for CRISPR
Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins

The following subsections go into more detail.

Sequencing a cell culture

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Diploid culture		buffer for ~1 week	days or more	(does tell euploidy)		diploid		mostly fine but accumulates mutation?	disrupts	
Haploid culture	a few papers; logically should work	? needs eggs; low-quality ok	days or more	(does tell euploidy)		haploid; combines well with number ID		mostly fine but accumulates a bit of mutation?	disrupts; unsure how much	

By far the most widely used chromosome identification method is sequencing (or SNP-array genotyping). You obtain some living cells with a known genotype by destructively sequencing some other cells; those other cells share a *common ancestor* with your current living cells.

Normally we wouldn't even think of this as an "indirect" genotyping method, but technically it's indirect in the sense of this article: it works by separating two items (the sample to be sequenced is removed from the cell culture), destructively sequencing one item (the sample), and then inferring the genotype of the other item (the cells that remain alive in the culture). It does blur the line between direct and indirect, though: If I give you a single cell, in order to sequence it non-destructively you must culture the cell (if you can), which does involve the cell changing (and therefore its genome potentially accumulating DNA or epigenomic damage).

Diploid culture

The basic method involves culturing ordinary diploid cells. Since this is a very standard method, I won't say much more, except to note:

- By strong default, culturing cells disrupts any reproduction-specific epigenomic state.
- Diploid sequencing would tell you that (most of) the cells in your culture are euploid diploid, and would tell you what homologs they contain. It would not tell you about any one specific chromosome. In other words, it's a strictly setwise identification method.
- Culturing in general has issues, which will be discussed in a future article. So, this method is not so appealing.
  - One of the main problems is that DNA tends to gradually accumulate damage in cultured cells.
  - One hypothetical way to alleviate this problem is "time travel sequencing". You take a sample from your culture, and then you literally *freeze* the rest of the culture. You sequence and analyze the sample, and prepare for the next operation you're going to apply to the culture. Only when you're ready do you thaw the culture. In this way, the cells don't divide and accumulate damage while you're doing other operations.
  - This method might not be workable; for example, freezing/thawing might stress the cells too much.

Haploid culture

See: Metacelsus (2025)[115].

Normally in a mammal's body, all cells are diploid, except for some specialized large tetraploid cells, and the haploid gametes. (Cancer cells might be aneuploid.)

Diploid cells can be cultured, but haploid cells don't naturally grow. This is especially true for gametes. The genome of a gamete is inactivated, so it doesn't produce the cell material needed to divide.

[115] Metacelsus. 'Androgenetic Haploid Selection'. Substack newsletter. De Novo, 16 November 2025. https://denovo.substack.com/p/androgenetic-haploid-selection.

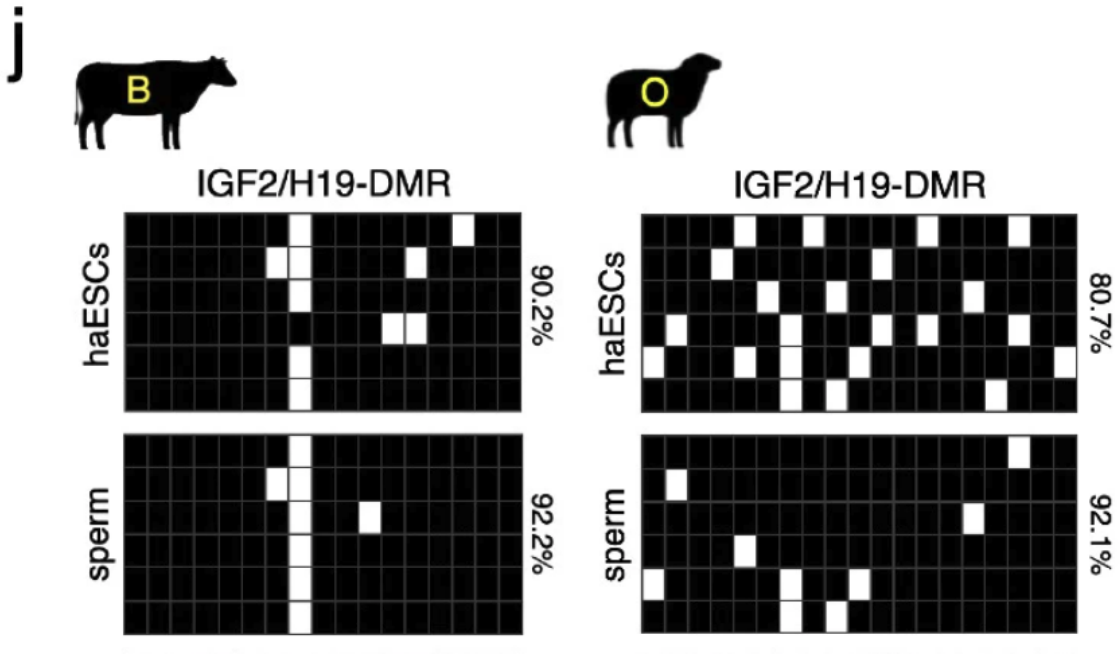
However, oocytes are very abnormally big cells, packed with those materials for growth. By activating an oocyte without a paternal haploid genome, Leeb and Wutz (2011) (as well as others) were able to grow cell lines with substantial portions of haploid cells for many weeks[116]. Similarly, Li et al. (2012) were able to grow *paternal* haploids transferred into oocytes, with the maternal DNA removed[117].

By culturing haploids, one could ("indirectly") sequence haploids. This gives setwise homolog identification.

This is an especially useful kind of setwise homolog identification, because it can, at least in principle, be used on sperm genomes. Since sperm genomes are the product of meiosis, they contain many different crossovers between the father's two chromosomes of each number. One could combine this method with chromosome-wise complementary identification on single spermatozoa. This would unlock recombinant chromosome selection.

Haploid culturing in general would very likely cause loss of imprinting, and generally disrupt the sperm epigenome. A possible way around this would be to culture the cells for only a short time. It may also be possible to prevent this loss, or mostly prevent the loss and then correct with epigenetic editing.

Yang et al. (2025) used this method to make haploid androgenetic cow stem cells, and then made cows from those haploids[118]. Most of the cows (and all of the sheep) either failed to grow or were stillborn or had developmental defects, but they did get one healthy fertile cow. The sperm epigenetic imprints that they checked looked fairly normal for the cow haploids (but not for the sheep ones):



(Extended Data Figure 5j from Yang et al. (2025). © publisher)

These are not good results for now, but are some kind of indication that this approach might work.

As Metacelsus points out: This would be a poor use of scarce oocytes today, but cheap oocyte-like cells might do the trick. They don't need to be correctly maternally imprinted or meiosed, they just need to provide the material for haploid growth for a few days. Hamazaki et al. (2021) were able to get stem cells to grow big like oocytes, and become able to support early-embryo-like growth[119].

[116]Leeb, Martin, and Anton Wutz. 'Derivation of Haploid Embryonic Stem Cells from Mouse Embryos'. Nature 479, no. 7371 (2011): 131–34. https://doi.org/10.1038/nature10448.

[117]Li, Wei, Ling Shuai, Haifeng Wan, et al. 'Androgenetic Haploid Embryonic Stem Cells Produce Live Transgenic Mice'. Nature 490, no. 7420 (2012): 407–11. https://doi.org/10.1038/nature11435.

[118]Yang, Lei, Anqi Di, Lishuang Song, et al. 'Generation of Modified Cows and Sheep from Spermatid-like Haploid Embryonic Stem Cells'. Nature Biotechnology, Nature Publishing Group, 7 October 2025, 1–9. https://doi.org/10.1038/s41587-025-02832-4.

[119]Hamazaki, Nobuhiko, Hirohisa Kyogoku, Hiromitsu Araki, et al. 'Reconstitution of the Oocyte Transcriptional Network with Transcription Factors'. Nature 589, no. 7841 (2021): 264–69. https://doi.org/10.1038/s41586-020-3027-9.

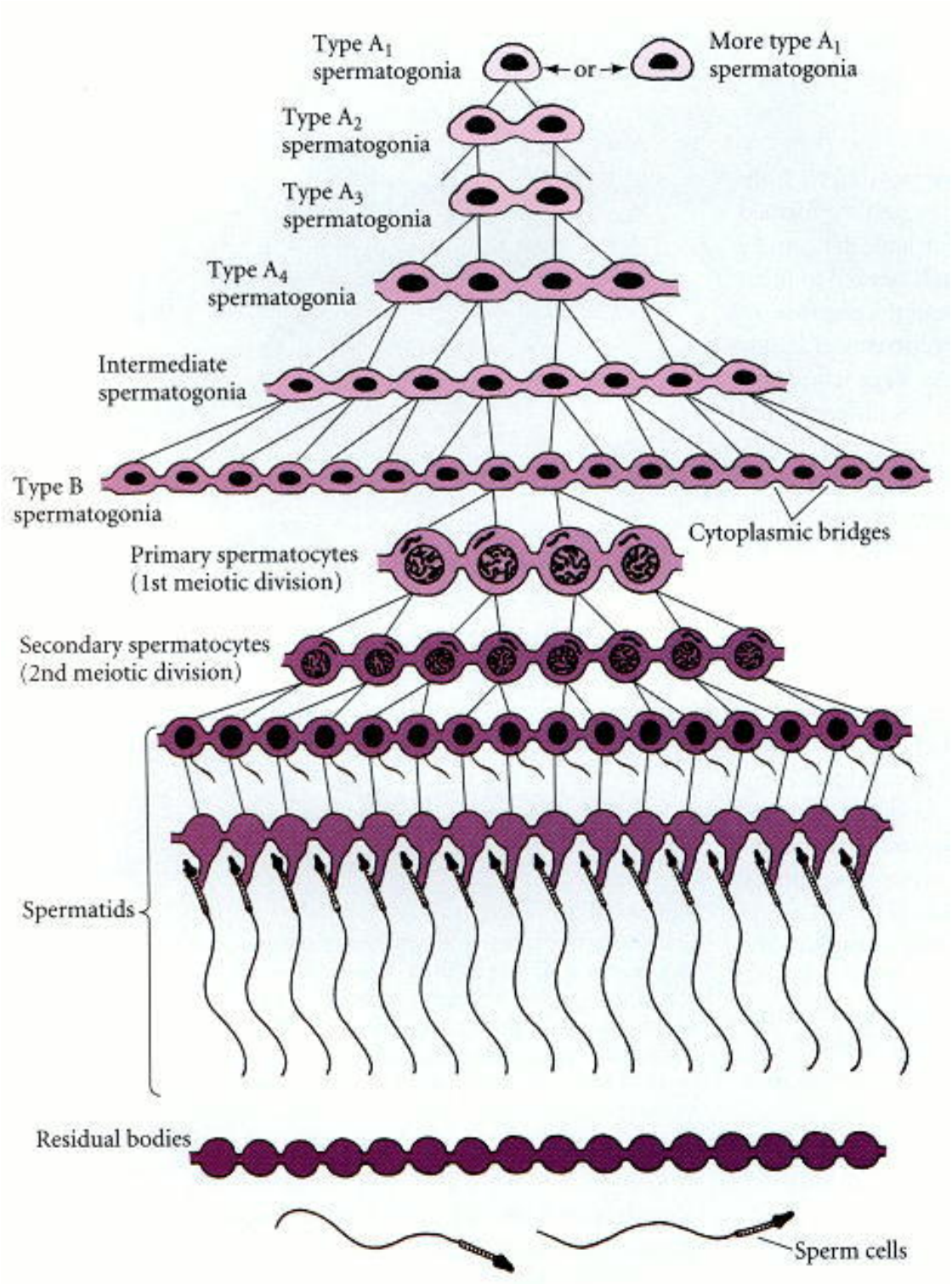
Sequencing post-meiotic RNA

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Sequencing post-meiotic RNA	1 paper; logically should work	? hard to develop	a day or so?	(does tell euploidy)		haploid; combines well with number ID				

Can we identify homologs setwise in gametes by looking at their RNAs?

The idea of RNA-based genotyping

Sperm develop from spermatogonial stem cells, which differentiate into diploid spermatocytes; the spermatocytes then proliferate and finally undergo meiosis:



(Figure 19.18 from Gilbert (2000)[120]. © publisher)

Until meiosis occurs, all the spermatogonia and spermatocytes have almost identical genomes (except for usually-small de novo mutations accumulating in the spermatogonial stem cell population). During meiosis, chromosomes recombine by crossover. After that point, different germline cells have different genomes. If two cells have different genomes, they would produce a slightly different profile of RNAs via transcription. Can we use the differences between the genomes of different germline cells to identify which crossovers occurred in the chromosomes of a given cell?

One might expect that spermatids (post-meiotic haploid cells) are transcriptionally silent. Since these cells are preparing to swim, they're shutting down most normal cellular processes and compacting their DNA. But, it turns out that they are definitely not transcriptionally silent. Up until partway through spermiogenesis (where spermatids become sperm), before the spermatids start elongating, there is a lot of transcription[121].

[120]Gilbert, Scott F. 'Spermatogenesis'. In Developmental Biology. 6th Edition. Sinauer Associates, 2000. https://www.ncbi.nlm.nih.gov/books/NBK10095/.

[121]Geremia, R., C. Boitani, M. Conti, and V. Monesi. 'RNA Synthesis in Spermatocytes and Spermatids and Preservation of Meiotic RNA during Spermiogenesis in the Mouse'. Cell Differentiation 5, no. 5 (1977): 343–55. https://doi.org/10.1016/0045-

Kierszenbaum and Tres (1978) state that "RNA synthesis occurs in the early steps of spermiogenesis but not in later"[122]. Eddy (2002) states[123]:

> Transcription occurs throughout germ cell development until the midpoint of the postmeiotic phase. During the latter half of this phase, proteins are synthesized from transcripts that have been stored since the early part of the postmeiotic phase.

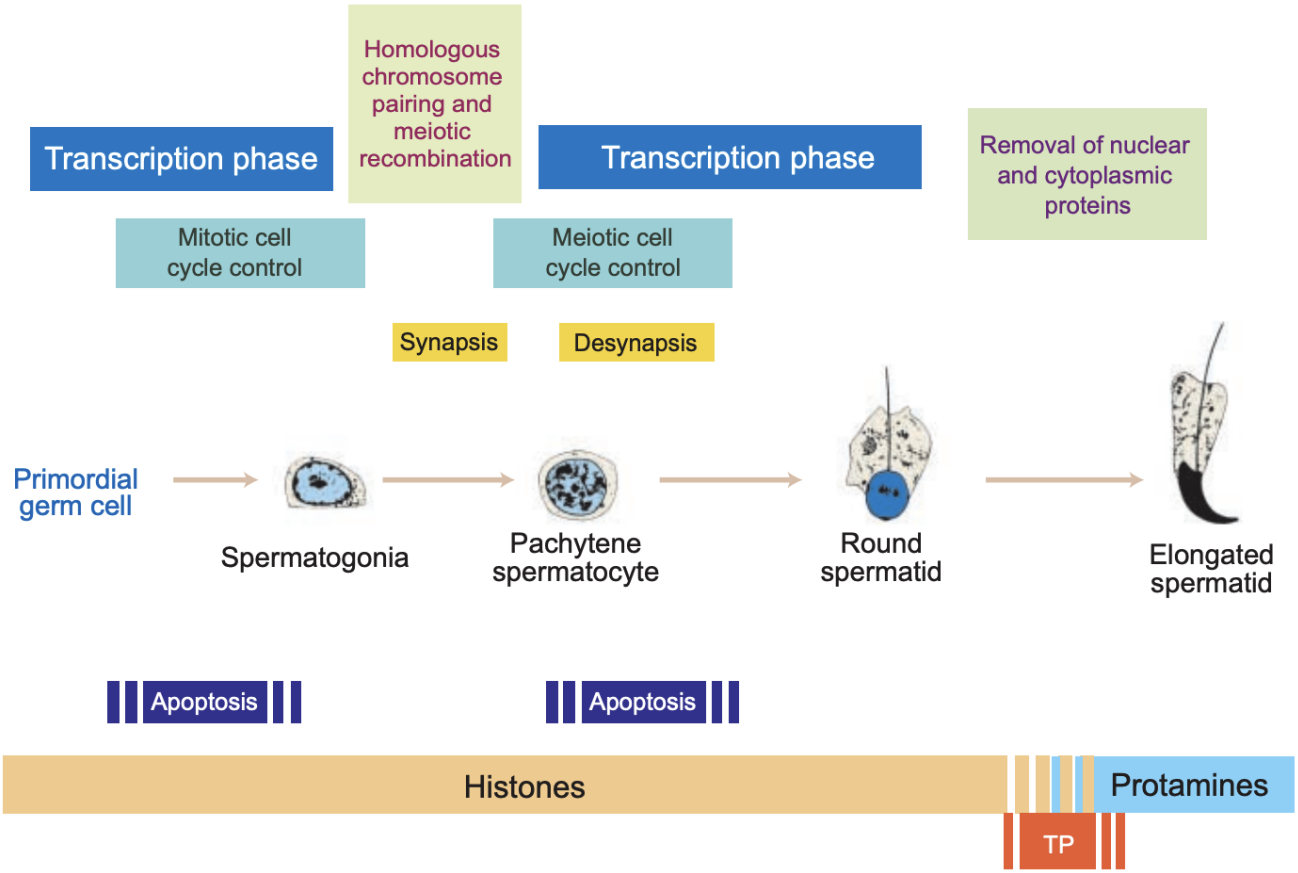


Fig. 1. Spermatogenesis is a cyclic developmental process by which spermatogonia cells generate the mature spermatozoon. These events are characterized by important modifications in chromatin organization, basically during two periods, meiosis—which includes the synapsis and desynapsis of the chromosomes—and the histone-protamine transition. Postmeiotically, a powerful wave of transcription occurs in haploid cells, which is governed by highly specialized molecular mechanisms. Specific genes operate at distinct steps of the spermatogenic process.

(Figure 1 from Sassone-Corsi (2002)[124]. © publisher)

The problem of spermatid RNA sharing

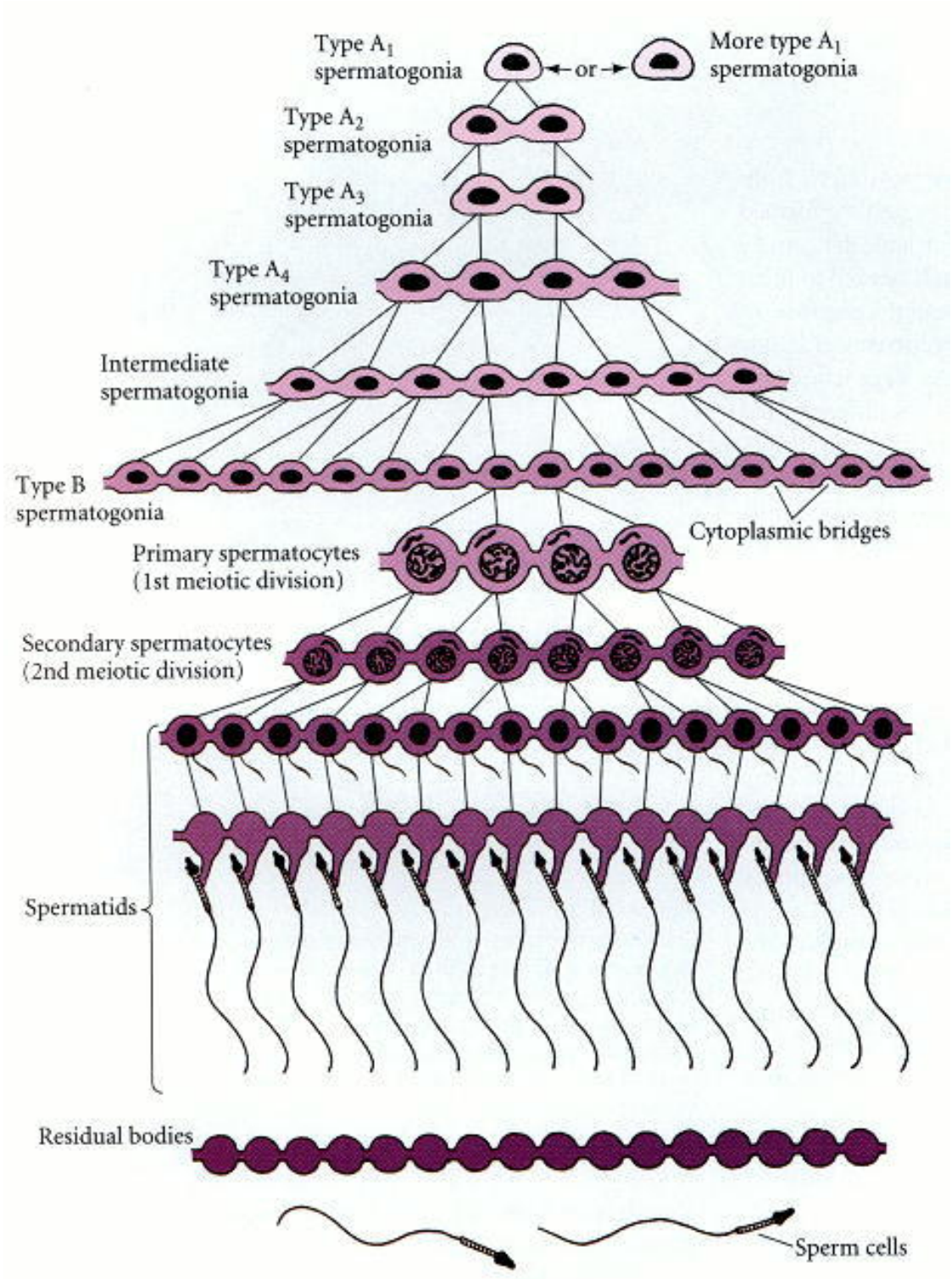
Does this imply that we can (setwise) identify homologs contained in a sperm, by sequencing the RNAs it contains? Not obviously. Recall the process of spermatogenesis:

6039(77)90072-0.

[122]Kierszenbaum, A L, and L L Tres. 'RNA Transcription and Chromatin Structure during Meiotic and Postmeiotic Stages of Spermatogenesis'. Federation Proceedings 37, no. 11 (1978): 2512–16. https://europepmc.org/article/med/357185.

[123]Eddy, Edward M. 'Male Germ Cell Gene Expression'. Recent Progress in Hormone Research 57 (2002): 103–28. https://doi.org/10.1210/rp.57.1.103.

[124]Sassone-Corsi, Paolo. 'Unique Chromatin Remodeling and Transcriptional Regulation in Spermatogenesis'. Science 296, no. 5576 (2002): 2176–78. https://doi.org/10.1126/science.1070963.



Note that in the later stages, where spermatids are growing into sperm, the spermatids are actually still connected to each other. When spermatocytes are differentiating and dividing in the preamble to meiosis, they don't fully divide; the mitotic and meiotic daughter cells stay connected to each other through cytoplasmic bridges. Therefore, we might expect that any RNAs being transcribed inside immediately post-meiotic spermatids will migrate between spermatids. (We'd especially expect spermatids to mix cytoplasm and RNAs with their nearest neighbors, which would be the other daughter cells of the primary spermatocyte they came from; i.e., the four spermatids containing four complementary haploid genomes, causing maximum confusion.)

RNA coverage of the genome

There's a further wrinkle, if we want to identify homologs in a sperm cell via RNA sequencing. We need to have many different RNAs transcribed from many different genomic loci, across most or all of the chromosomes. If there are many chromosomes without at least a couple RNAs being transcribed, then we can't infer much about what crossovers those chromosomes have. In particular, we'd need there to be many RNAs transcribed from loci where the maternal and paternal DNA *differ* (in the person producing the sperm).

As one piece of data, Hermann et al. (2018) seem to say that there is a large degree of transcription in spermatids[125]. If I'm reading it correctly, the sheet "Human_mouse_std.comparison" in their table S4 (copied here: https://docs.google.com/spreadsheets/d/1VH3MJRhhwjZmPEZKZ4Q8yva26oX4GH16oV4928Obbn0/) seems to list hundreds of genes that are noticeably expressed in human spermatids, and the first 20 or so seem to be from a random scattering of chromosomes.

Would there be enough SNPs between maternal and paternal chromosomes? Probably, going off of SNP density. The normal density is about 8 SNPs / 10kb. But even within exons, the density is still around 5 SNPs / 10kb. And within introns, which make up about 10x as much of the genome as exons, the density is also 8 SNPs / 10kb[126]. The upshot is that the strong default guess would be that there are plenty of available SNPs to use as tags for identifying crossover locations.

(We could also consider sequencing proteins, rather than RNAs. I don't know if that would be better somehow; probably RNA sequencing is better in most ways, but for example, proteins are more abundant (at least in normal cells). However, this would substantially restrict the genomic loci that we're measuring. Also, we might get an inaccurate or blurred measurement, because proteins present in a spermatozoon might have substantially come from translated RNAs that were produced before meiosis (and therefore aren't distinctive to an individual sperm). Finally, proteins, being smaller than many RNAs, may be more prone to cross cytoplasmic bridges, thus mixing between spermatids.)

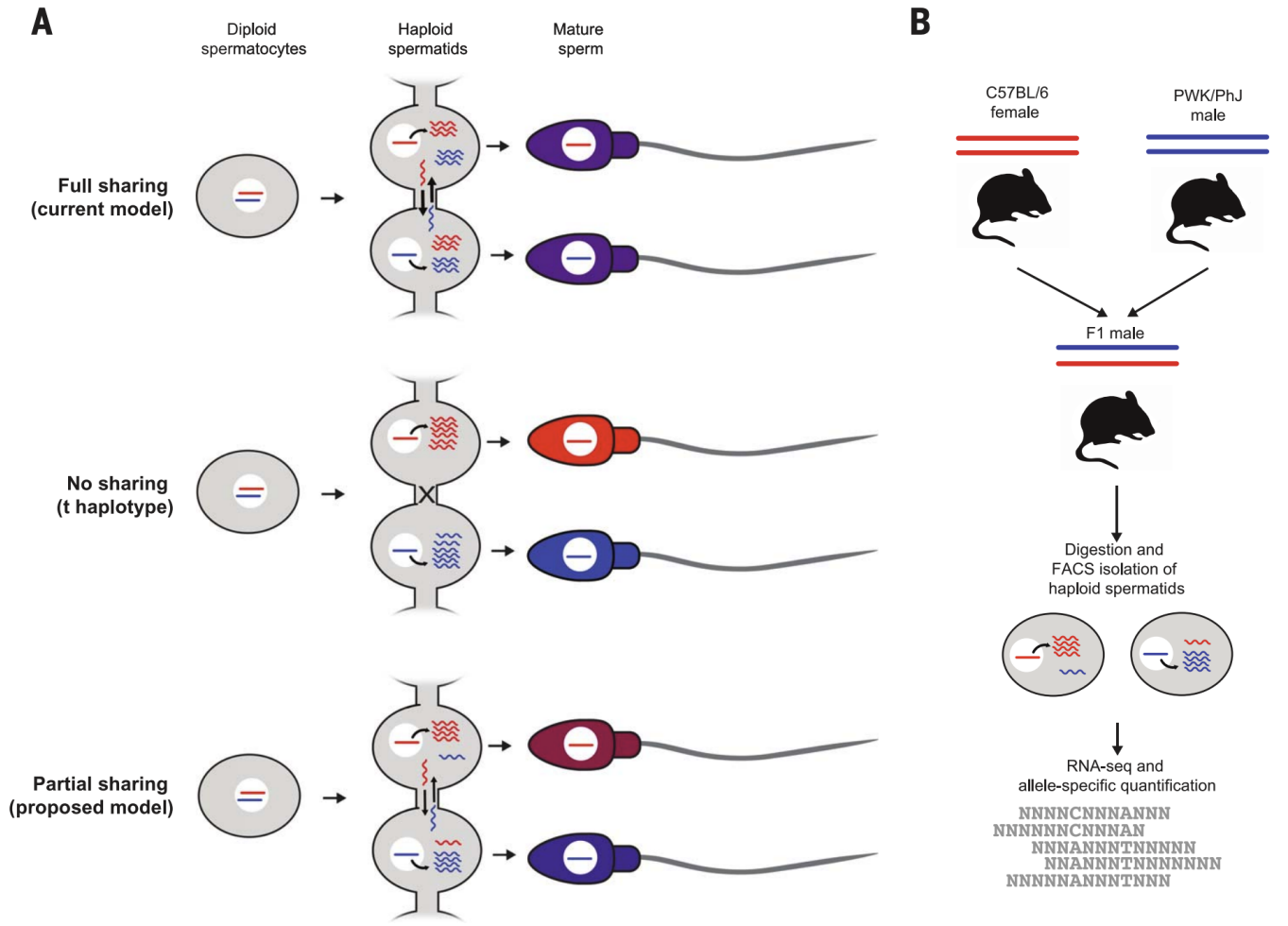
Proof of concept

As it turns out, empirically, for whatever reason, different spermatids do in fact have distinctive profiles of RNAs. This was investigated by Bhutani et al. (2021)[127]. They studied spermatids in mice, a bull, two cynomolgus macaques, and two humans, sequencing the RNAs found in spermatids at different stages. They found that very many (thousands) of homologous RNAs were expressed in a way that was heavily biased towards one or the other allele. This, and other evidence, led them to suggest that a large subset (roughly a third) of RNAs expressed during the early spermatid stages will partially or completely avoid crossing cytoplasmic bridges:

[125]Hermann, Brian P., Keren Cheng, Anukriti Singh, et al. 'The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids'. Cell Reports 25, no. 6 (2018): 1650-1667.e8. https://doi.org/10.1016/j.celrep.2018.10.026.

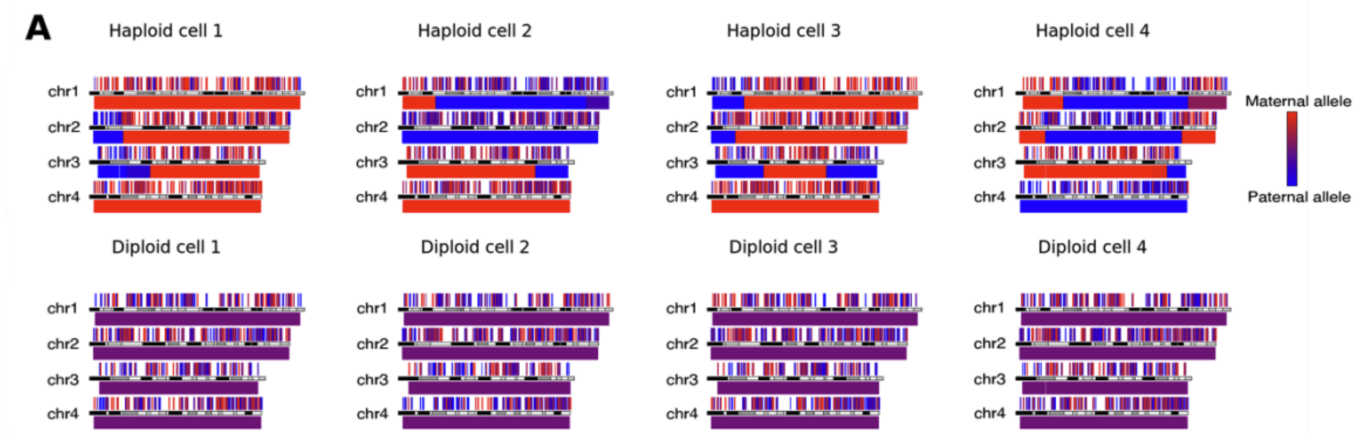
[126]Zhao, Zhongming, Yun-Xin Fu, David Hewett-Emmett, and Eric Boerwinkle. 'Investigating Single Nucleotide Polymorphism (SNP) Density in the Human Genome and Its Implications for Molecular Evolution'. Gene 312 (July 2003): 207–13. https://doi.org/10.1016/S0378-1119(03)00670-X.

[127]Bhutani, Kunal, Katherine Stansifer, Simina Ticau, et al. 'Widespread Haploid-Biased Gene Expression Enables Sperm-Level Natural Selection'. Science 371, no. 6533 (2021): eabb1723. https://doi.org/10.1126/science.abb1723.



(Figure 1A,B from Bhutani (2021). © publisher)

In particular, by looking at the overall profile of allele-biased RNA expression, they could reconstruct what crossovers likely occurred (in mice):



(Figure S2A from Bhutani (2021). © publisher)

Note that, as I understand it, they didn't actually sequence the DNA of these specific haploid spermatids. But, at least for the mice, they had phased chromosomes of the male producing spermatids (by breeding two different homozygous strains); so they can verify the fact that the inferred chromosomes look like the two paternal chromosomes plus 0-2 crossovers. That's good evidence that they are indirectly measuring homologs.

They had problems with phasing the human DNA (more homozygosity, fewer SNPs). So it's not directly verified that this could work well in humans. But, by more extensively sequencing the father, one could fully phase his chromosomes, and there should be plenty of SNPs, so this ought to work.

The hard part would be dissociating the RNAs from the chromatin without destroying the chromatin.

An additional challenge would be dealing with the epigenomic state of the spermatid DNA. During late spermiogenesis, most histones in spermatid DNA are replaced with protamines, and this is likely to be

epigenetically important. In theory, spermatids could be matured in vitro to complete that process. However, compared to spermatids, mature spermatozoa are even more compact and might have significantly less RNA available to sequence. There's some indication this might still work—see Tomoiaga et al. (2020)[128].

RNA-based genotyping for oocytes?

Would this work with oocytes, rather than sperm? It's quite unlikely. There's several blockers:

- GV oocytes seem to be already transcriptionally deactivated, and will remain so until after fertilization[129]. GV oocytes are still in meiosis I; in particular, they still have a diploid (copied) genome. So, logically, we wouldn't expect to find any allele-biased RNA expression in oocytes.
- Even if there were transcription at later stages, even fully mature (ovulated) oocytes still have 46 chromosomes, as they're arrested in meiosis II. If they had transcriptional activity, it would be biased, so you would be able to infer some of the homology of the final haploid genome of that oocyte. But on average the effect would be lessened (I'm not sure how much, it might depend on the pattern of crossovers in oocytes in general). In theory you could sequence the haploid polar body expelled by the oocyte after fertilization, which ought to complete your identification of the homologs in the final haploid oocyte genome; but this is getting complicated.
- Oocytes are very expensive and precious, and these methods would be destructive. So any method for chromosome selection on oocytes is probably too costly and risky for most people.

If in vitro oogenesis were achieved, these constraints might be lifted. In particular, in theory one could artificially activate an oocyte, and then sequence the RNAs in that oocyte to infer the homologs in that oocyte. However, this might be much harder than just doing haploid culturing.

Complementary identification

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Chromoso complemer identificati	not shown; logically should work	? hard to develop	minutes?		combines with set-wise homolog ID	1				
Meiotic complemer identificati	1 paper; logically should work	medium unit price??	days?	(does tell euploidy)		haploid; combines well with number ID	much less confident with < 3 meiotic cousins		(creates it; or else disrupts)	(creates them)

(See the earlier subsection "Destructive identification and complementary identification".)

In general, the idea of complementary sequencing is:

1. You have multiple sets $S_1, ..., S_k$ of chromosomes.
2. You know the full genotype (homolog or number), as a set, of the union $\bigcup_i S_i$ of those sets.
3. You directly measure the genotype (homolog or number) of all but one of the sets, say $S_2, ..., S_k$.
4. You indirectly infer the genotype of the remaining set as $S_1 = \bigcup_i S_i \setminus \bigcup_{i \neq 1} S_i$.

The following subsections give several variants of complementary identification, depending on where the sets S_i come from.

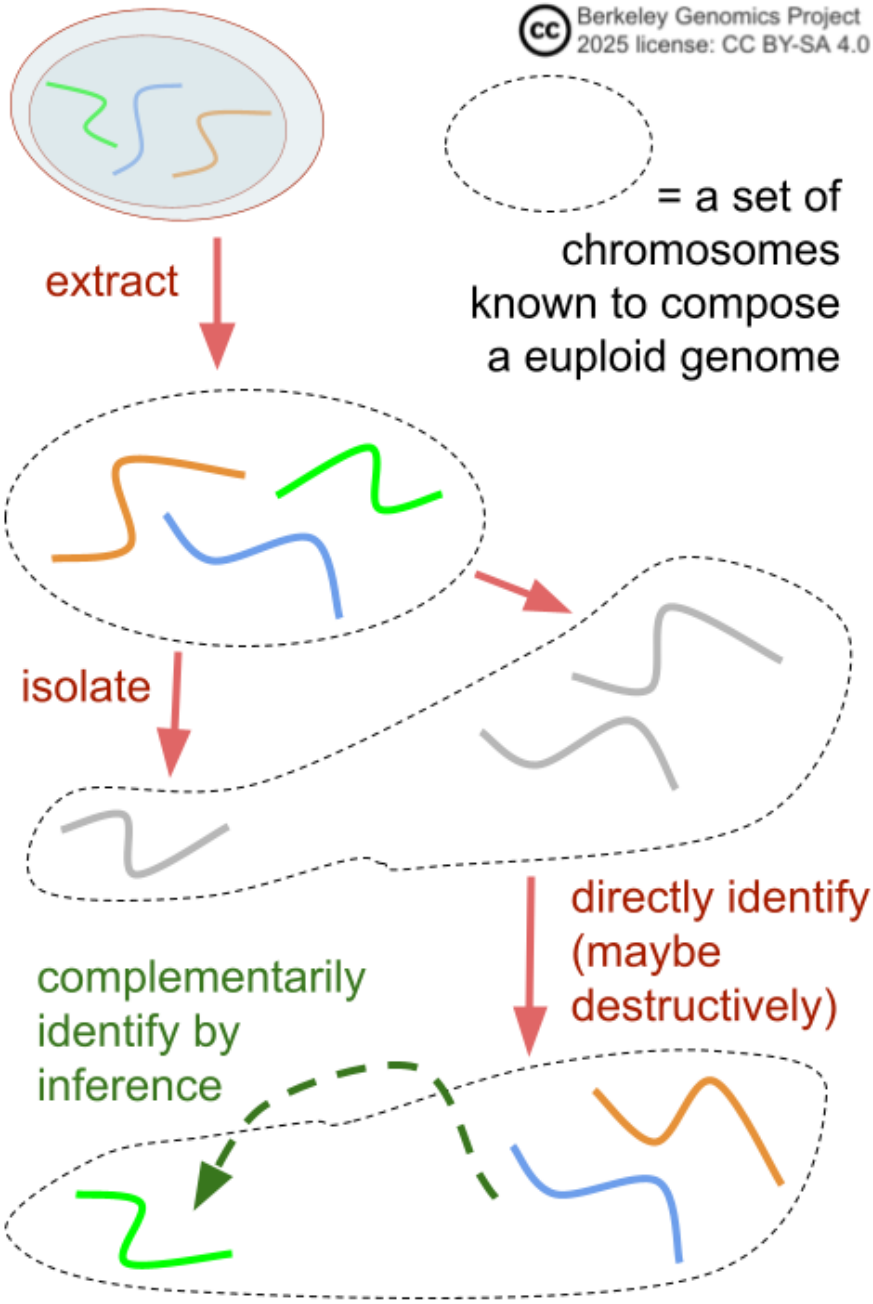
[128]Tomoiaga, Delia, Vanessa Aguiar-Pulido, Shristi Shrestha, et al. "Single-Cell Sperm Transcriptomes and Variants from Fathers of Children with and without Autism Spectrum Disorder." Npj Genomic Medicine 5, no. 1 (2020): 14. https://doi.org/10.1038/s41525-020-0117-4.

[129]Bogolyubova, Irina, Daniil Salimov, and Dmitry Bogolyubov. 'Chromatin Configuration in Diplotene Mouse and Human Oocytes during the Period of Transcriptional Activity Extinction'. International Journal of Molecular Sciences 24, no. 14 (2023): 11517. https://doi.org/10.3390/ijms241411517.

Chromosome-wise complementary identification

In chromosome-wise complementary identification, we have a single cell with a known set of chromosomes. We could either have the set of chromosome *numbers* be known, or else have also the chromosome *homologs* be known.

Then we genotype all but one of the chromosomes, identifying them (by number or by homolog). There should be one chromosome missing from the genotype that we obtain. That missing chromosome is the one that we didn't genotype, and instead preserved separately. Thus we infer the genotype of the separated chromosome. This can also be done by holding out a set of chromosomes, rather than just one; then we infer what set of chromosomes is held out (but not which one is which).



If we knew the set of numbers of the full starting set, and we identify the numbers of the non-preserved set, then we infer the set of numbers of the preserved set. If we also knew the homologs present in the starting set (e.g. because we sequenced the parent or the cell line), and we identify the homologs in the non-preserved set, then we infer the set of homologs in the preserved set.

We could know the starting set by some other method (e.g. sequencing a cell culture). We might also just presumptively know. E.g. we might presume that an ordinary cell is diploid euploid, or we might presume that a gamete is haploid euploid.

If the homolog is already determined by the starting set, then we don't need to measure the homologs. For example, suppose you have a diploid cell that is homozygous at chromosome number 1. If you do complementary chromosome number identification to get a chromosome 1 from such a cell, then you already

know what homolog you have.

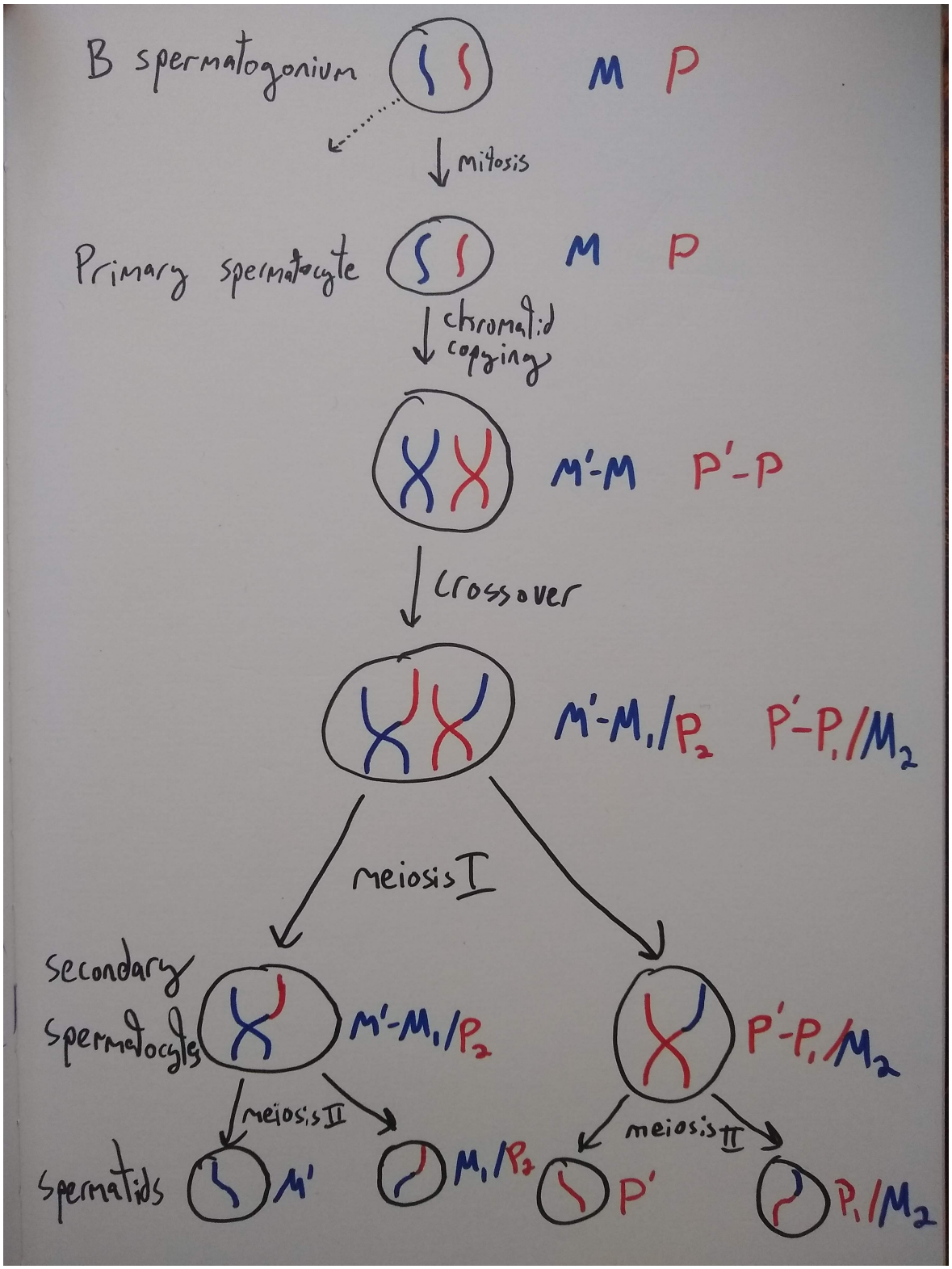
The implementation of chromosome-wise complementary identification will be discussed in future work. As a teaser trailer, see Babahosseini et al. (2021)[130].

Meiotic complementary identification

This is a method for setwise homolog identification in haploid cells.

In meiotic complementary identification, we have a diploid grandparent cell. That cell undergoes meiosis, producing four haploid granddaughter cells. First it undergoes meiosis I, copying its DNA and crossing over chromatids and then dividing into two daughter cells, which are haploid but have two connected copies of each chromosome. Second, it undergoes meiosis II, separating the connected copies, producing the four granddaughter cells which are haploid, with one copy of each chromosome:

[130]Babahosseini, Hesam, Darawalee Wangsa, Mani Pabba, Thomas Ried, Tom Misteli, and Don L DeVoe. "Deterministic Assembly of Chromosome Ensembles in a Programmable Membrane Trap Array." Biofabrication 13, no. 4 (2021): 10.1088/1758-5090/ac1258. https://doi.org/10.1088/1758-5090/ac1258.



(Figure from Benson-Tilsen (2022)[131]. This figure has some inaccuracies but communicates the basic structure of meiosis.)

The idea of meiotic complementary identification is to capture all four haploid meiotic granddaughter cells that come from a single diploid grandparent cell. We know that collectively, these four haploids have a double diploid genome. In other words, they'll contain four chromosome number 1s—two copies of each of the two chromosome number 1s that were in the diploid grandparent cell.

Then we sequence three of the four haploids. By complementation, we can then infer the set of DNA that's in the remaining haploid cell. Specifically, we know the homologs present in that remaining cell—we know where the crossovers happened. And we didn't have to directly interact with that haploid cell.

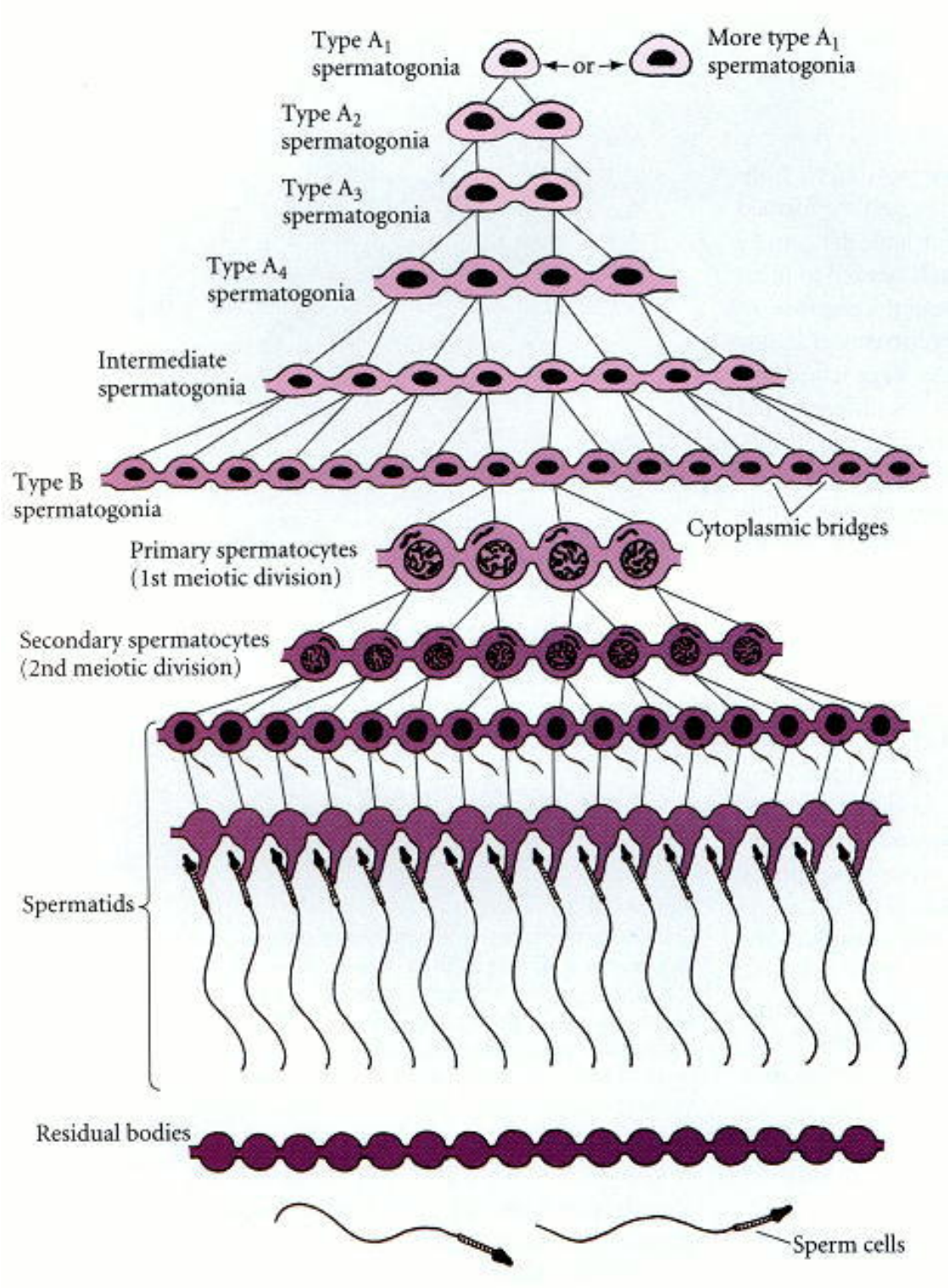
So, meiotic complementary identification would give a setwise, non-destructive, indirect homolog identification method. Is this workable in practice?

[131]Benson-Tilsen, Tsvi. 'Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins'. Non-Destructively Sequencing Gametes by Sequencing Meiotic Cousins, 29 June 2022. https://tsvibt.blogspot.com/2022/06/non-destructively-sequencing-gametes-by.html.

In general, if you have an in vitro meiosis method that works on individual cells, then plausibly you could do meiotic complementation. You'd just have to keep each diploid and its granddaughter haploids separate from each other.

For normal natural spermatogenesis, it's not workable. Spermatogenesis is complete before gametes leave the testes, so the common origin is not available. (In theory you can do a really complicated expensive scheme where you do single-cell sequencing to millions of sperm to identify triplets with an implied fourth meiotic cousin sperm, and then sequence sperm chromosomes one at a time to notice when a sperm is a fourth meiotic cousin—but this is not practical.)

Even if we could do in vitro spermatogenesis, which is not yet feasible in humans, complementary identification might be infeasible. That's because the spermatocytes that are about to undergo meiosis are not individual cells:

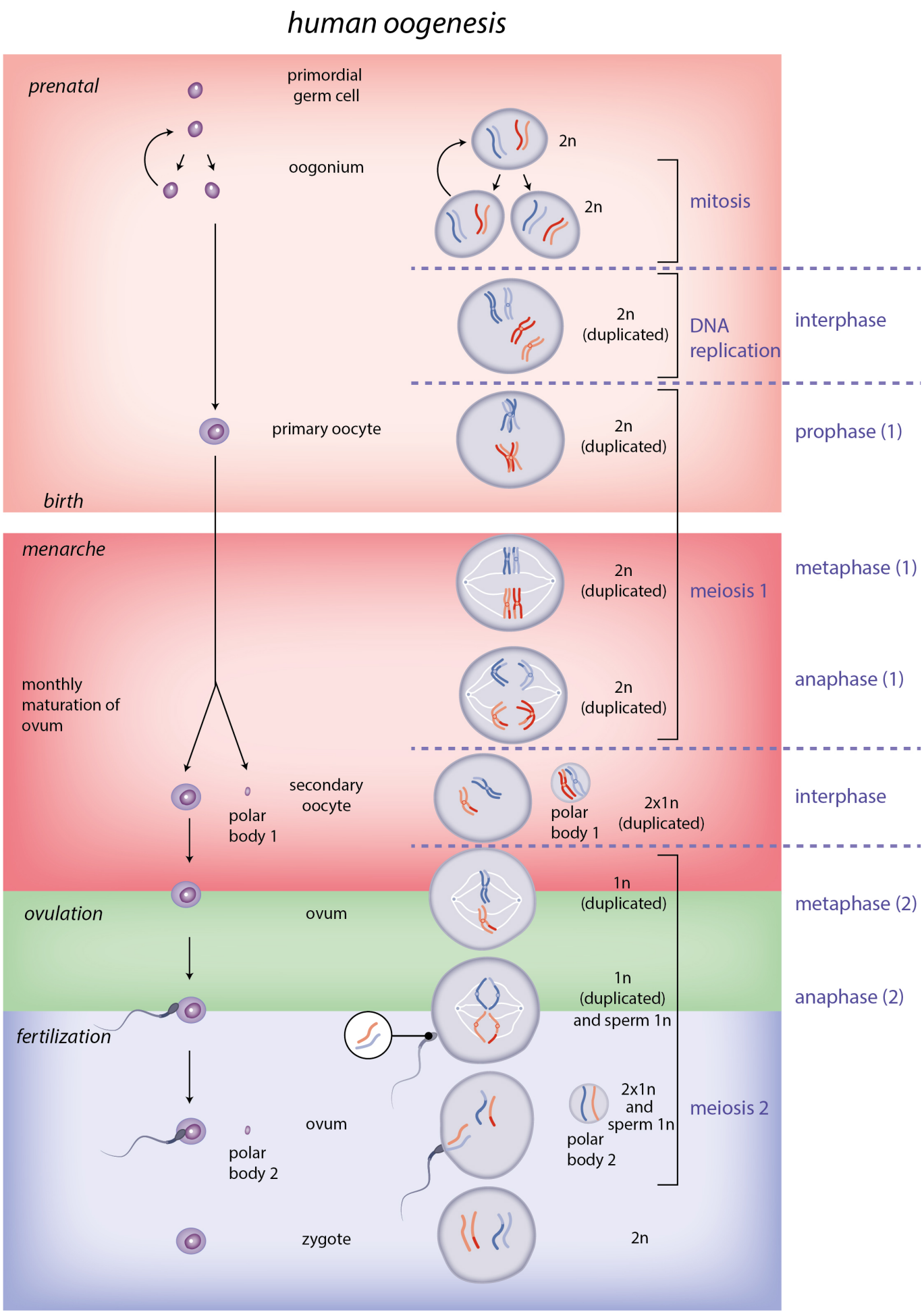


(Figure 19.18 from Gilbert (2000)[132]. © publisher)

[132]Gilbert, Scott F. 'Spermatogenesis'. In Developmental Biology. 6th Edition. Sinauer Associates, 2000. https://www.ncbi.nlm.nih.gov/books/NBK10095/.

Instead, spermatocytes stay connected to each other through cytoplasmic bridges, well into the final stages of spermatogenesis, even after meiosis. So, at least by default, meiotic cousin sperm are not cordoned off as a set. But, an in vitro spermatogenesis protocol might produce separated primary spermatocytes, in which case complementary identification might be feasible.

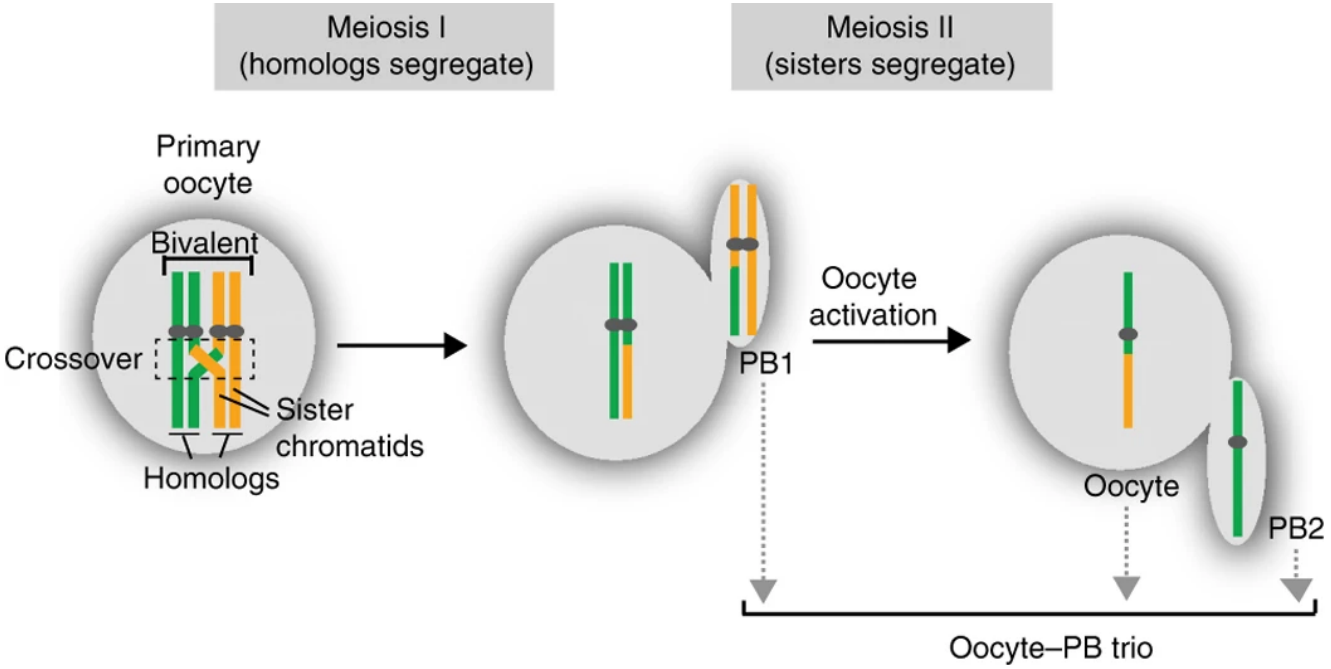
What about for oogenesis? Meiotic complementation might actually work for oogenesis. Here's a diagram of oogenesis:



(Diagram from Slagter et al.[133]. © publisher)

[133]Ron Slagter, O. Paul Gobée, LUMC, Hope Wicks, LUMC, et al. 'Slagter - Drawing Human Oogenesis Diagram - English Labels | AnatomyTOOL'. Accessed 21 February 2025. https://anatomytool.org/content/slagter-drawing-human-oogenesis-diagram-english-labels.

The key is that the first meiotic daughter cell, polar body 1, is often carried along with the oocyte through ovulation. If you recover polar body 1, and also recover polar body 2 later, then you have the full complement to the oocyte's haploid genome. Indeed, Ottolini et al. (2016) used this to infer the genotypes of oocytes[134]:



(Figure 1 from Ottolini et al. (2016). © publisher)

It's probably not that feasible to do chromosome selection on natural oocytes anyway, simply because oocytes are scarce. In vitro oogenesis should enable meiotic complementation.

However, in both natural and in vitro oogenesis, there's a problem with the second polar body: it doesn't exist until after fertilization; before fertilization, the oocyte genome still has 2 copies of each chromosome fused together. This means that to retrieve polar body 2, you'd have to artificially activate the oocyte so that it completes meiosis II. That's doable, but it's an additional complication, and it would risk affecting the epigenomic state of the oocyte genome.

It might be unnecessary to retrieve polar body 2. We would already partially infer the homolog set of an oocyte's haploid genome just given polar body 1; we'd resolve about half of the uncertainty (I think). That would provide much of the benefit of homolog identification.

I don't know whether meiotic complementation would be easier than, say, haploid culturing.

Targeted DNA operations

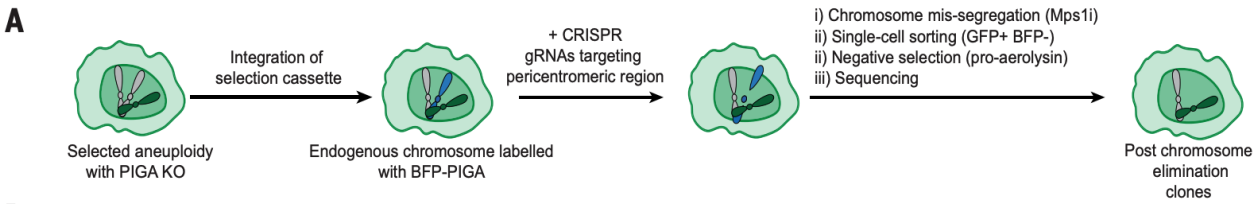
There's a class of methods for manipulating chromosomes that could be called identification methods. These methods involve performing some operation on some specific chromosome, targeted by number or by homolog. Examples:

- Digestion / elimination / cleavage / destruction.
  - Get rid of a specified chromosome.
  - For example, a CRISPR system could be used to target a chromosome number or an allele, inducing a double-strand break at that locus. Depending on context, one or more double-strand breaks would tend to permanently destroy the chromosome.
- Selection.
  - Make cells with a specified non-desired chromosome grow slower, and/or make cells with a specified desired chromosome grow faster.
- Missegregation.
  - Make a specified chromosome be more likely to fail to separate from its copy during mitosis.
  - Some of the resulting daughter cells have two copies of that chromosome, and some have zero. These cells can then be selected, e.g. based on ploidy.

[134]Ottolini, Christian S., Antonio Capalbo, Louise Newnham, et al. 'Generation of Meiomaps of Genome-Wide Recombination and Chromosome Segregation in Human Oocytes'. Nature Protocols 11, no. 7 (2016): 1229–43. https://doi.org/10.1038/nprot.2016.075.

There are many targeting methods—methods for attaching to specific DNA regions, and what to attach. I haven't studied them. In addition to the methods listed in "Hybridization methods", there are magnetic beads, fluorophores, etc., and zinc finger proteins (using much more gentle fixation?), transcription activator-like effectors, etc.

For example, Petris et al. (2025) were able to selectively eliminate one of three homologous chromosomes from cells that received a transplanted third chromosome[135]:



(Figure 4 from Petris et al. (2025). © publisher)

These methods blur the line between indirect and direct identification. For example, selection could be done by *promoting* cells that contain chromosomes that include a specific marker (direct interaction with desired chromosomes); or by *demoting* those cells (indirectly promoting the prevalence of desired chromosomes); or both at the same time.

These methods also blur the line between chromosome identification and active chromosome manipulation. For example, chromosome elimination increases the relative frequency of the desired chromosome within the targeted sample, thus indirectly identifying the remaining chromosomes as more likely to be the desired ones. (Compare bulk sorting, which works similarly.)

Finally, most of these methods are for contexts where we're culturing cells.

For these reasons, I'll defer some more discussion of these methods to future work.

Targeted complementary elimination

Method	Refined	Price	Time	Identifies numbers	Identifies homologs	Set size	Confident	Saves DNA	Saves epigenetic state	Saves structural proteins
Targeted complemer elimination	?	?	hours?		with more design work?	1		probably ok?	? may require fixation like for CRISPR	? may require fixation like for CRISPR

Targeted chromosome elimination is a possible exception to the rule that targeted DNA operations happen in culturing contexts.

We could isolate the chromosomes from a single cell, and then eliminate all chromosomes except for the desired one using targeted elimination. We could then isolate the remaining desired chromosome.

One of the issues with targeted elimination in a culturing context is that DNA damage response mechanisms in cells would tend to kill the cell if there's too much damage. But if the chromosomes have been isolated, those mechanisms would not be tripped. So, many chromosomes could be eliminated at once.

This could be simpler than complementary identification in at least one respect: it should only require a single aliquot, so it should avoid the need for complex microfluidics. However, the targeted elimination method in full would have to avoid damaging the desired chromosome. Also, the systems used to eliminate the chromosomes add more complexity, and are likely to require a significant degree of protein digestion to get access to the DNA.

Acknowledgements

Thanks to Ben Korpan for related discussions. Thanks to supporters of the Berkeley Genomics Project.

[135]Petris, Gianluca, Simona Grazioli, Linda van Bijsterveldt, et al. 'High-Fidelity Human Chromosome Transfer and Elimination'. Science 390, no. 6777 (2025): 1038–43. https://doi.org/10.1126/science.adv9797.