DNA Fingerprinting

During the last class we discussed DNA self-assembly

It is perhaps the ultimate self-assembly process

Wanted to study it for inspiration and because of potential non-biological application

But having gone that far, I just had to discuss its best known application: DNA fingerprinting

DNA technology IS nano, but it's also a huge field usually taught in its own courses

That temptation extends to my decision to add a DNA fingerprinting lab to this class

In which we will attempt to identify "Student X" by a process known as "VNTR PCR"

The companion manual for that lab can be downloaded at:

https://WeCanFigureThisOut.org/NANO/labs/materials/UVA_DNA_fingerprinting_manual.pdf

Sources:

As a non-molecular biologist, in preparing this lecture I have consulted texts including:

Molecular Biology of the Cell by Bruce Alberts

The World of the Cell by W.M. Becker, J.B. Reece & M.F. Poenie

Biochemistry by D. Voet and J. Voet

Which I will credit in figures as: “Alberts,” "World of the Cell" or "Voet & Voet"

I have also made use of DNA education websites

Particularly that of of the Dolan DNA Learning Center at Cold Spring Harbor Laboratory

And, finally, I have reviewed online lecture note sets prepared by other professors

There are two common types of DNA fingerprinting:

RFLP or "R-FLIP"

Was invented first (~1990)

Requires very little background knowledge of genomics

But has disadvantages of requiring:

- Large tissue samples
- Un-degraded samples
- Use of radioactive tracers

STR/VNTR PCR

Technique NOW used

Eliminates all of RFLP's disadvantages above

But requires significant background information and terminology from genomics

DNA fingerprinting based on

Restriction Fragment Length Polymorphism

(RFLP)

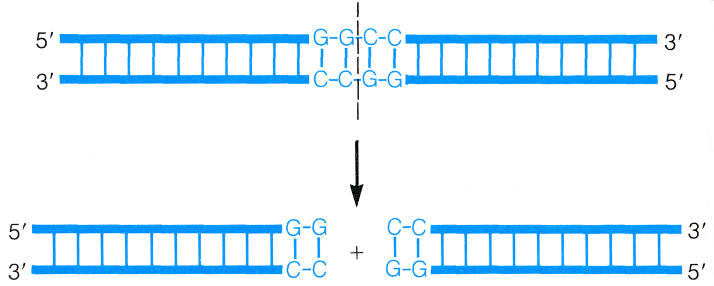
Based on a really clever trick bacteria developed for fighting off viral attacks:

You've probably heard that viruses commandeer cell mechanisms to replicate their DNA

Some bacteria combat this by never using a certain short base sequence in their own DNA

Then create a **restriction enzyme** that will cut any DNA at that sequence

For instance, "HaeIII" separates DNA in middle of GC GC CG CG sequence (only!):



(from World of the Cell figure 14B-1)

Where virus uses this sequence => its DNA is immediately sliced and diced!

Question: Isn't **never** using a certain four base sequence a big handicap?

No: DNA uses a three base sequence ("codon") to specify an amino acid => protein

(4 choices) x (4 choices) x (4 choices) = 64 codes, but life only uses 20 different amino acids

So have at least three available codes for each amino acid, so giving up one = No problem!

Such a simple base sequence also likely in human genome:

In fact, would expect any four base sequence to occur MANY times in 1 meter of our DNA

But exceedingly unlikely to occur at identical places in DNA of two individuals

Providing basis for RFLP DNA fingerprinting:

1) Treat human DNA with such a restriction enzyme

DNA will be cut at every point where restriction sequence occurs

Creating **restriction fragments**

2) Separate various length segments produced (a.k.a. **length polymorphisms**)

3) Compare differences in fragment lengths between individuals

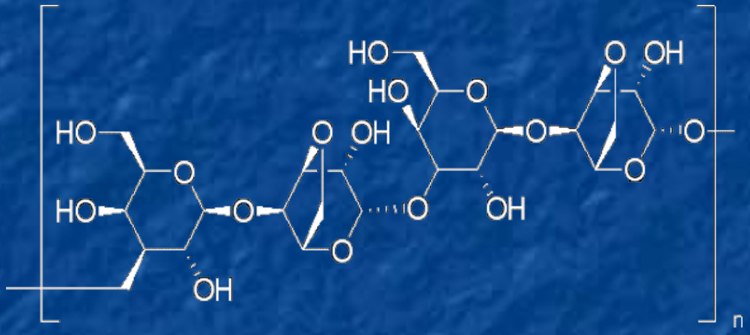
How to separate different length fragments of DNA?

Gel Electrophoresis:

Start by making a “gel” = Mesh of organic polymers + Water

Most commonly used polymer is agarose (derived from seaweed):

Repeating units of:



Makes a mesh through which DNA fragments can migrate

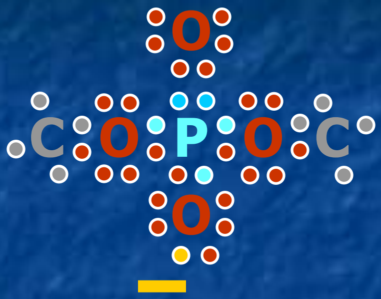
Shorter DNA fragments can move more easily through this mesh

Migration rates can be changed by altering agarose concentration

More agarose => smaller pores => slower migration

But then <u>drive</u> migration by applying an electric field:

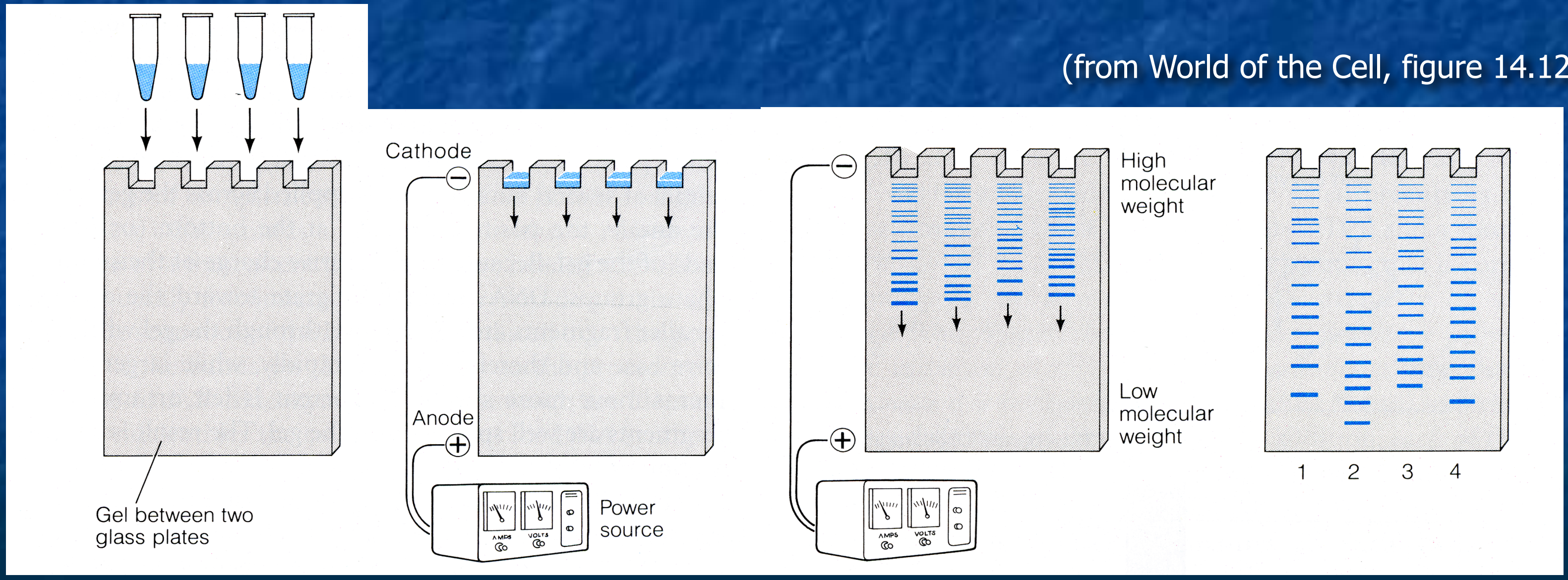
Exploit fact that charged phosphate backbone units give DNA fragments net negative charge:



Insert DNA segments into channels, positive voltage then draws them through gel

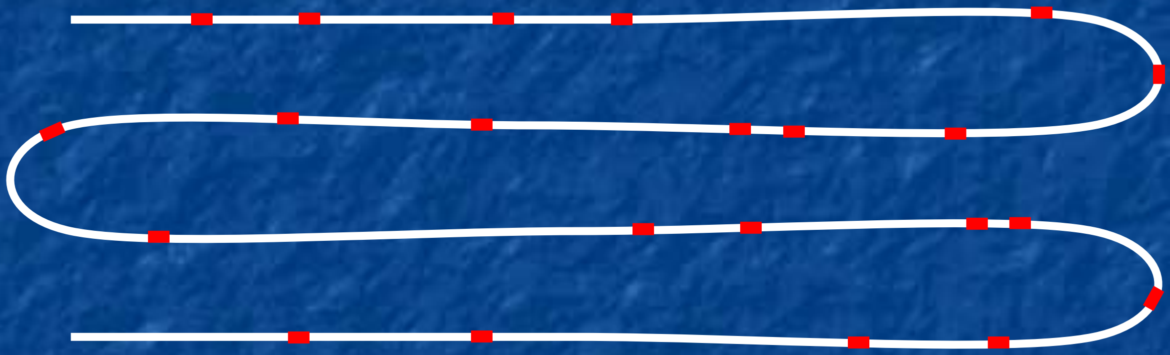
With shorter DNA fragments racing ahead:

(from World of the Cell, figure 14.12)

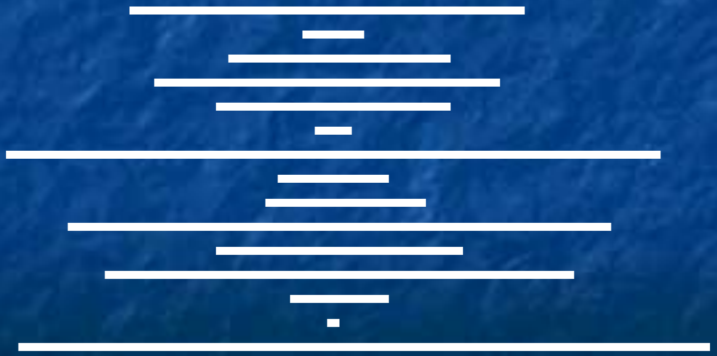


Would work fine with moderate number of restriction sites

Representation of entire length of genome (with selected restriction sites indicated in red):



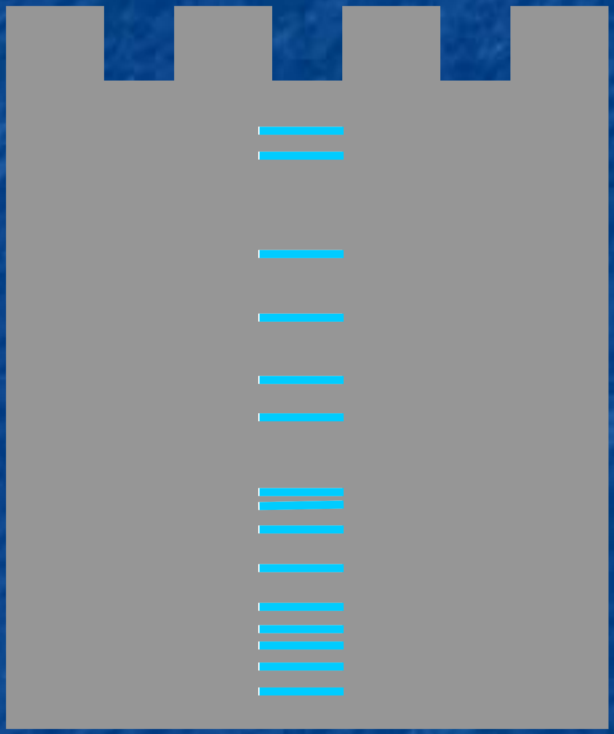
Yielding DNA restriction fragment length assortment something like this:



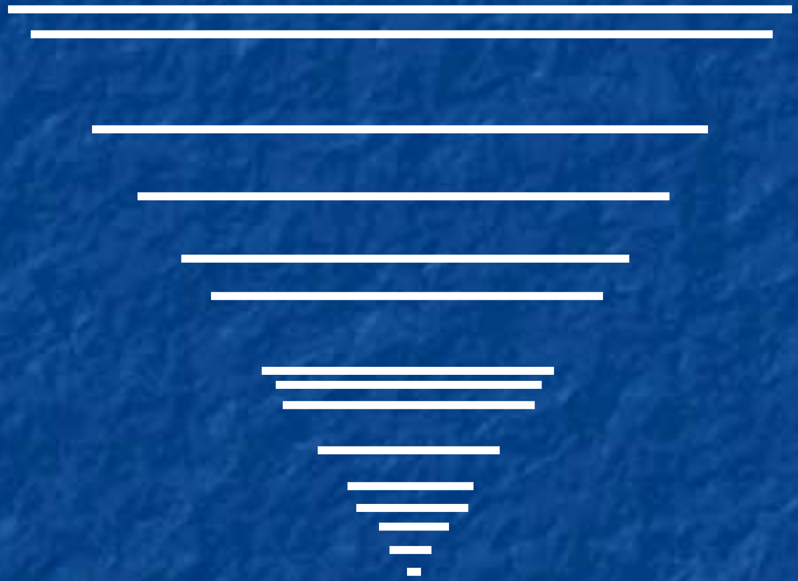
A Hands-on Introduction to Nanoscience: WeCanFigureThisOut.org/NANO/Nano_home.htm

After electrophoresis and staining:

Electrophoresis gel channel



Corresponding DNA fragments

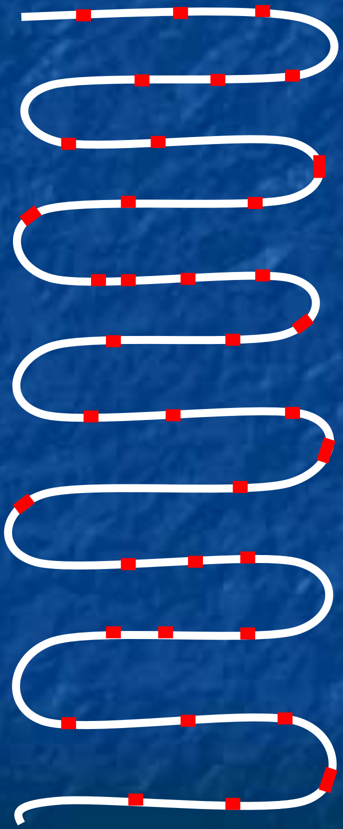


Looks like this will work just fine, what's the problem?

In fact are likely to be huge number of restriction sites!

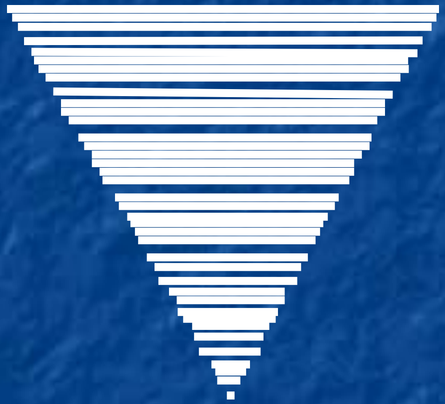
So slightly more realistically:

DNA:

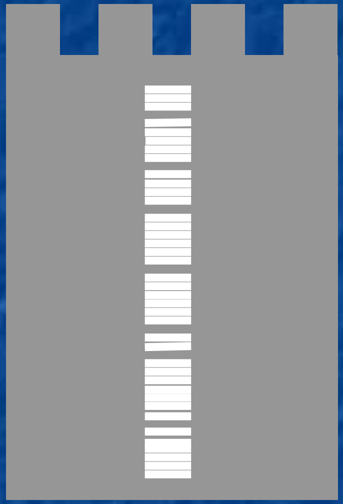


(and more)

Restriction fragments



Electrophoresis Gel:



Probable result: Virtually continuous array/smear of fragment sizes

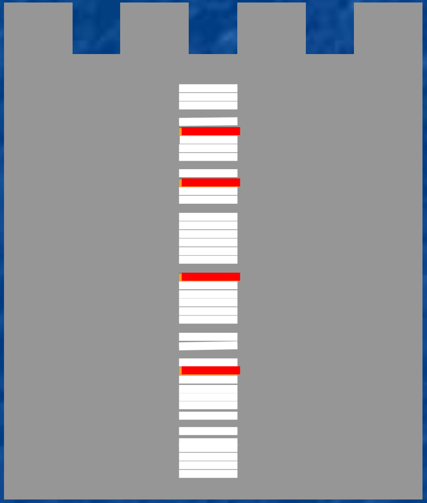
Loosing ability to differentiate between samples

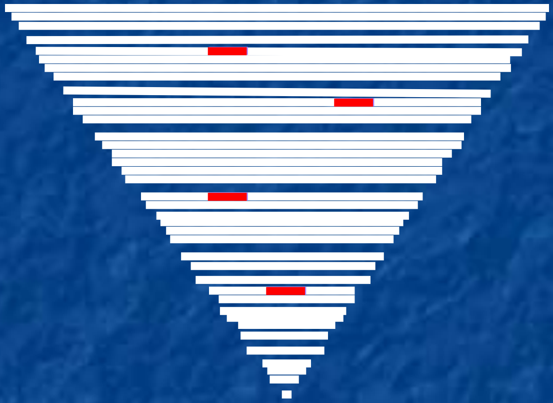
Rescue process by adding one more step:

Blot the above gel with a predetermined reference segment of DNA: 🟧

That **probe** DNA will combine with complementary portions (🟪) of sample’s DNA => 🟧🟪

With proper selection/complexity of probe DNA, will only match in rare test fragments:





Add radioactive atoms to probe DNA => Its stripes in gel will expose photographic film

Shortcomings of RFLP Fingerprinting?

Need a fairly large initial test sample

Even with blotting, need many copies of any fragment to get detectable signal

And there is no replication ("amplification") of original sample in this technique

Need "undegraded" sample

If sample is degraded, and some strands lost, signal strength problem (above) worsens

If degradation includes breaking of strands (which is likely), it will produce:

New fragments NOT DUE TO RESTRICTION SITES

Adding new FALSE lines to fingerprint!!

Leading to the alternative modern fingerprinting technique:

DNA fingerprinting based on

Polymerase Chain Reaction

(PCR)

Advantages of PCR fingerprinting:

Can analyze much smaller samples

Less sensitive to sample degradation

Requires no radioactively tagged probes

Disadvantages of PCR fingerprinting:

Process is significantly more complex

Understanding requires deeper scientific background information

To begin with that necessary background information:

Similarity & dissimilarity of human DNA

1) Between different humans, DNA is almost identical:

Overlap believed to be as high as 99.8%

2) Large fraction (~98.5%) of DNA not known to code creation of RNA or proteins:

Was once classified as “junk” = non-functional DNA baggage left over from evolution

Supported by fact that some simpler organisms (e.g. lily) have much more DNA

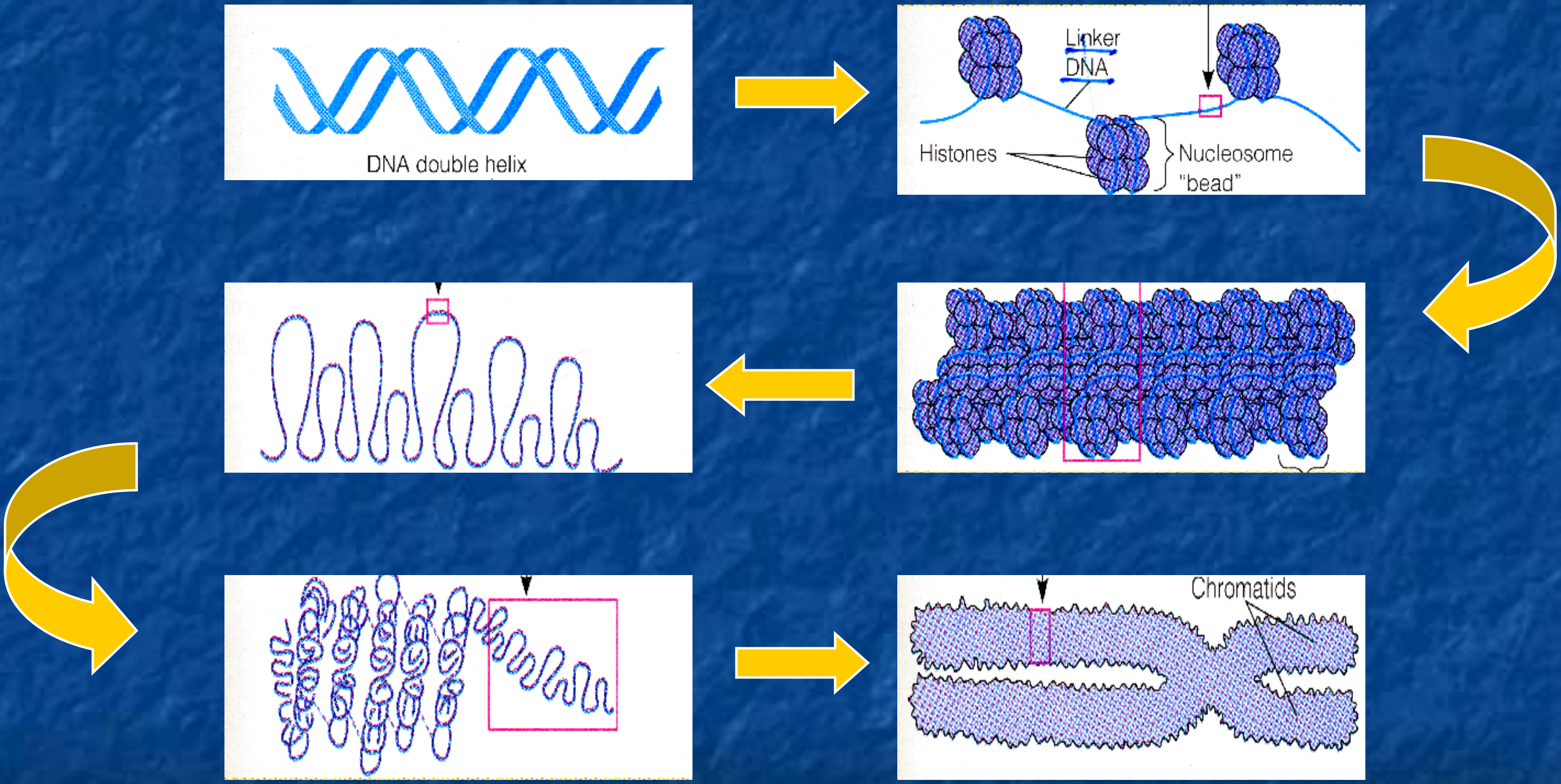
But fraction labeled “Junk” being nibbled downward as our understanding increases:

For instance, in emerging understanding of “epigenetic” control of genetic expression

3) Are variations between individuals in BOTH functional & non-functional DNA

Organization of genome (at least during cell division)

In preceding class, we saw how ~ 1 meter of DNA was wrapped up inside cell nucleus:



(from World of the Cell figure 14.8)

Relevance (or irrelevance) to PCR DNA fingerprinting:

Fingerprinting almost ignores organization and treats DNA as a continuous 1 meter length

Exception comes at the final chromosome level of organization:

A specific “gene” (DNA section programming specific inheritable characteristic)

occurs at specific location, in specific chromosome

(Biologists love Latin, so location => **locus**, locations => **loci**)

But, from high school, recall that we have we have pairs of chromosomes:

One containing DNA from mother, one containing DNA from father

So, depending on whether given gene from mother & father is identical or not

We have either one or two variations of every gene in our genome

Leading to a bit more necessary terminology:

Term applied to variants (or versions) of a given gene = **alleles**

If versions of gene from mother & father are identical, have one “allele” of that gene

Identical “alleles” = **homozygous**

If versions are different, then you have two “alleles” of that gene

Different “alleles” = **heterozygous**

But the term **allele** is also applied to non-gene sections of DNA

For any locus on genome, can talk about variations (alleles) between chromosome pairs

With implications:

In a **single individual**, for a given gene / locus on genome:

Individual has **1 allele** (DNA at that locus on chromosome pairs is identical)

OR

Individual has **2 alleles** (DNA at that locus on chromosome pairs is different)

In **population**, for given gene / locus on genome:

There can be (and almost certainly will be) **MANY alleles**

Multiplicity of variations at specific site is key to PCR fingerprinting technique

But this leads to a relevant question:

If each of our parents had two examples of every gene (identical or non-identical)

How does child of two parents avoid having four examples?

Reduction process occurs in formation of reproductive ovum and sperm cells (“gametes”):

Normal cells have two examples of every gene / DNA segment = “diploid”

Reproductive cells have one example of every gene / DNA segment = “haploid”

Process by which diploid cellular DNA is reduced to haploid gamete DNA is called

Recombination: Process by which two DNA strands => one hybrid strand

It’s a mix and match process with many variations:

Inner workings of one DNA recombination process:

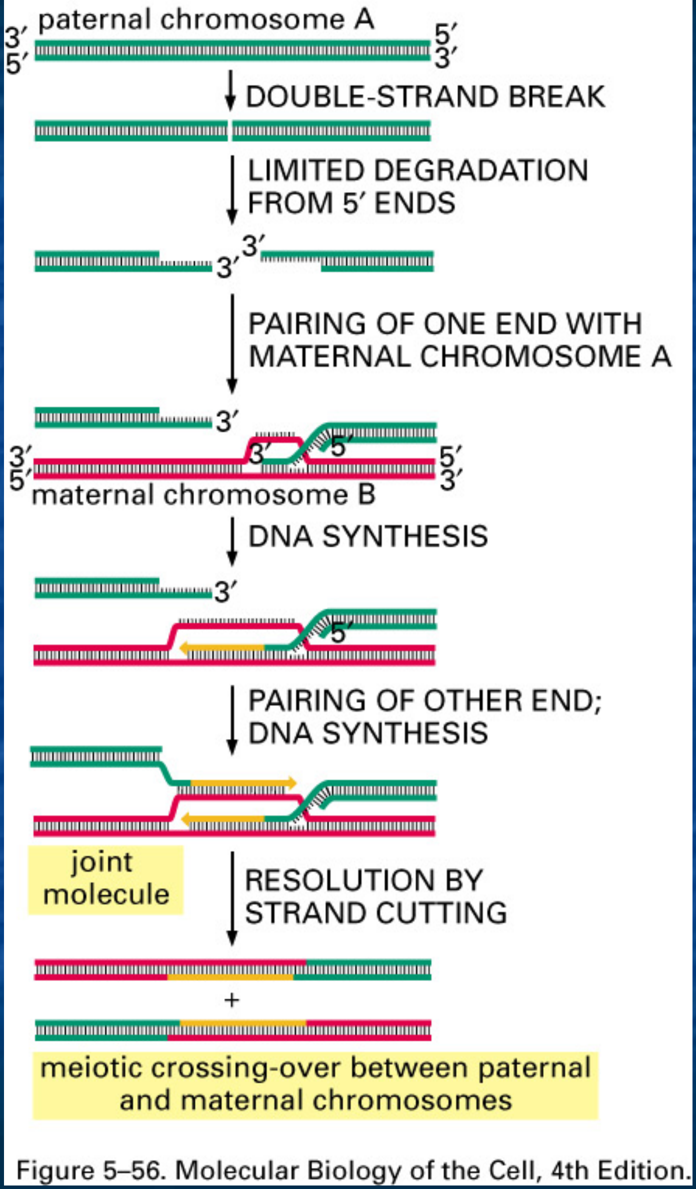

paternal chromosome A
3′
5′
5′
3′
DOUBLE-STRAND BREAK
LIMITED DEGRADATION FROM 5′ ENDS
3′
3′
PAIRING OF ONE END WITH MATERNAL CHROMOSOME A
3′
3′
5′
3′
5′
5′
3′
maternal chromosome B
DNA SYNTHESIS
3′
5′
PAIRING OF OTHER END; DNA SYNTHESIS
joint molecule
RESOLUTION BY STRAND CUTTING
+
meiotic crossing-over between paternal and maternal chromosomes

Figure 5–56. Molecular Biology of the Cell, 4th Edition.

Paternal DNA helix (green) cut at one point

Strands partially decompose, revealing base coding

Where matches, end of one strand can insert itself into maternal helix. Then:

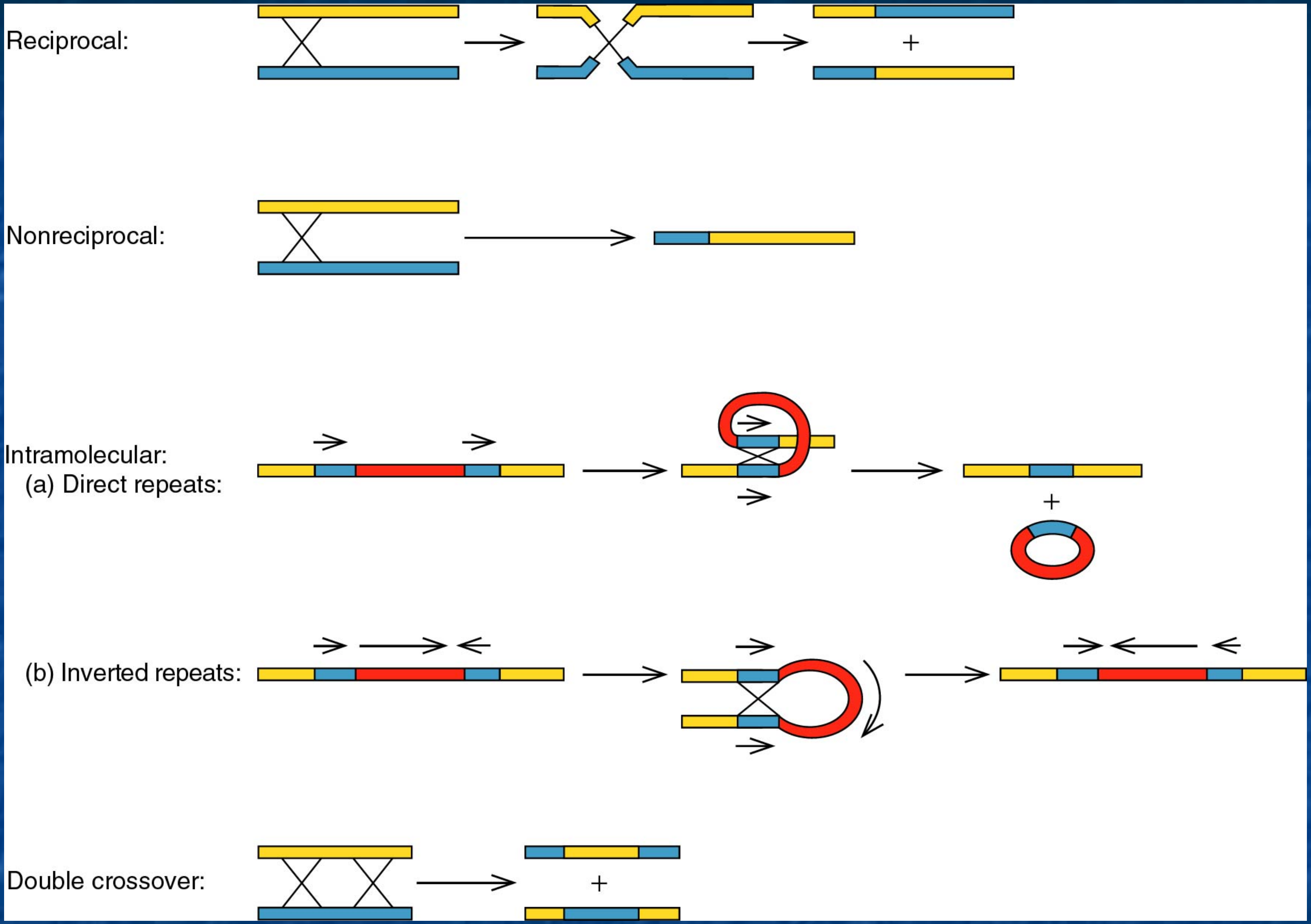
1) Inserted segment can extend on its 3’ end, further driving maternal strands apart

2) “Orphaned” maternal segment can then join with other complementary segment of other paternal strand

Ultimately yielding two helices, each containing sections of paternal and maternal DNA

Note: For this to work well, has to start in regions where helices are almost identical: Can thus begin in the middle of a protein encoding sequence (i.e. a gene)!

Plus even weirder schemes:



From online lecture notes entitled “Introduction of DNA Recombination” by Haoran Zhang, Department of Chemical & Biological Engineering Tufts University

A Hands-on Introduction to Nanoscience: WeCanFigureThisOut.org/NANO/Nano_home.htm

Relevance to PCR DNA fingerprinting?

"Mistakes" can be made during the above recombination processes:

Genes OR arbitrary sections of DNA can be duplicated on a given strand

Genes or DNA sections can be shifted from normal location to new location (locus)

These "mistakes" may actually offer an evolutionary advantage

Simple mutation of a gene is a gamble: May help, may hurt

But if gene is first duplicated, and one copy then mutates, still have a backup copy

So if that gene a created critical enzyme, that enzyme will still be produced!

This process may also have opened the door to epigenetics:

By providing genome with "alternate recipes" applicable in different environments

There is evidence that ENTIRE length of genome was duplicated once or twice!

Time out: Does this all begin to sound a little disorganized?

From the Albert text on molecular cell biology (p. 206):

The human genome seems to be in an alarming state of disarray. As one commentator described our genome:

"In some ways it may resemble your garage / bedroom / refrigerator / life:

highly individualistic, but unkempt;

little evidence of organization;

much accumulated clutter (referred to by the uninitiated as 'junk');

virtually nothing ever discarded;

and the few patently valuable items indiscriminatingly, apparently carelessly, scattered throughout."

So, accepting this haphazard organization and moving on:

Impact on DNA structure:

"Mistakes" in protein encoding (or controlling) DNA may yield non-viable embryo

So most of these errors will not propagate

Protein encoding/controlling DNA is thus "strongly conserved"

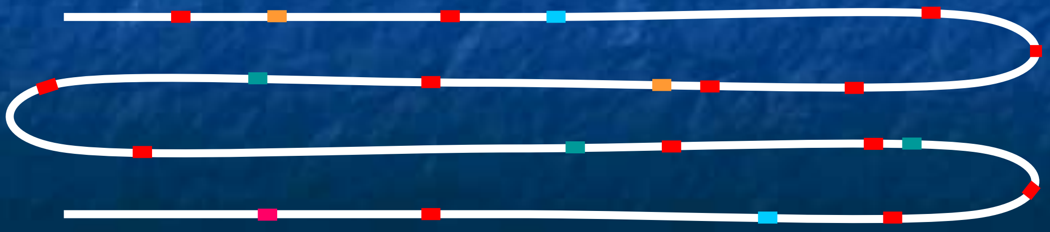
And subject to the least variation between individuals

But errors in non-protein encoding/controlling regions may cause no problems

So the these errors can accumulate leading to stretches of DNA called:

Markers = Points (loci) on the genome with well-known repetitive structures

Within a given marker, structure is identical or strongly similar, between individuals



Types of markers?

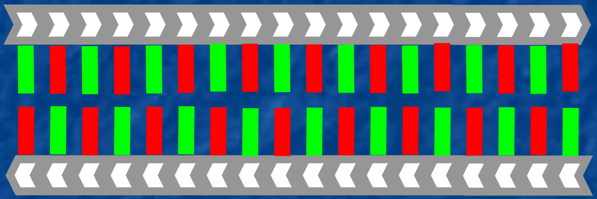
Many different types

But for PCR-based fingerprinting, most important are tandem repeats

Short base sequences repeated over and over again ("repeats")

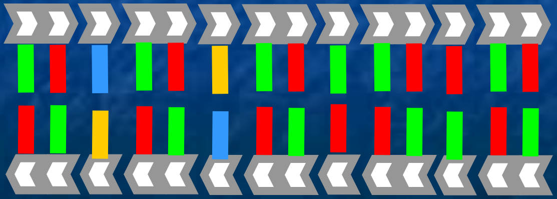
With no other intervening bases ("tandem")

So this would be a tandem repeat:



Called **STR's** = short tandem repeats OR **VNTR's** = variable number tandem repeats

But this would be a non-tandem repeat:



Differences & similarities of tandem repeat markers:

Because of “errors” in replication processes such as recombination,

at given tandem repeat marker, individuals can differ in number of repeats:

For example: Individual #1: $(AG)^N$ Individual #2: $(AG)^M$

At each tandem repeat marker, may be ~10-50 variations (alleles) in population as a whole

Use number of repeats at a given marker to differentiate individuals

If do this at 13 different tandem repeat marker locations (loci)

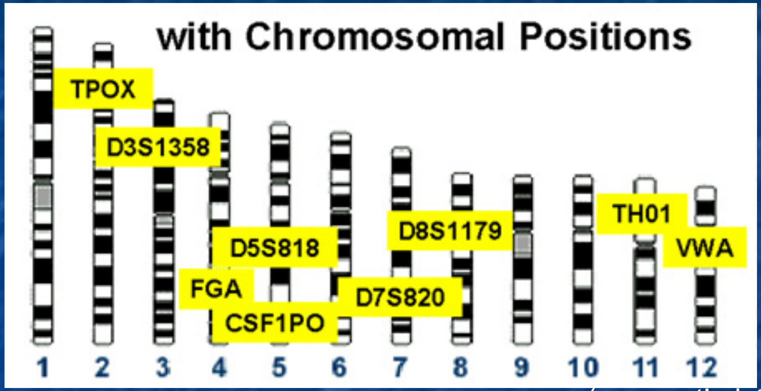
Product of variations at each marker (10-50) yields huge total range of possibilities:

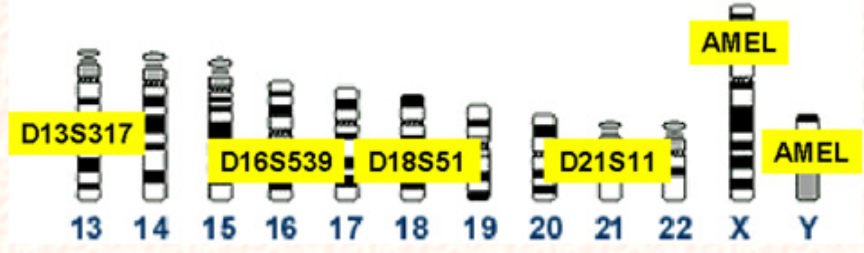
$10^{13} - 50^{13} = 10^{13} - 10^{22}$ => VERY unlikely that two individuals have same allele set

This is the basis of the FBI’s CODIS DNA fingerprinting database

13 "Loci" recorded by FBI "Combined DNA Information System" (CODIS)

All come from locations on one of the 23 "chromosome" pairs of DNA held in nuclei of cells





(www.cstl.nist.gov/strbase/fbicore.htm)

Nuclear DNA is used in CODIS because it **differs the most** between individuals

But other times want DNA that **differs the least** between **related** individuals

Thus "National Missing Person DNA Database" (NMPDD) instead uses:

- Loci on Y chromosome - inherited only from father

- Mitochondrial (non-nuclear) DNA – inherited only from mother

Mitochondrial DNA also more concentrated in cells => easier to detect

Problem:

How do I locate and extract the tandem repeat sequences AT the 13 CODIS loci?

I need to make sure I get material from all 13 (and no material from elsewhere)

Also need to make sure I get FULL tandem repeat at each locus

Because if I randomly shorten the number of repeats (by cutting out only part),

I am changing the content of the fingerprint

Nature gives us a break:

Number of repeats differs between individuals

But those repeat sequences can be **surrounded** by DNA that is identical in all individuals

So can use FIXED surrounding DNA to locate target STR/VNTR segments

Targeting tandem repeat segments using the polymerase chain reaction:

Polymerase chain reaction is a variation of the natural DNA replication mechanism:

- Mix bath with DNA, soup of different nucleotides, primers, and enzyme (**polymerase**)

- Heat to ~ 95 C: DNA double strands separate (heat overcomes base pair hydrogen bonds)

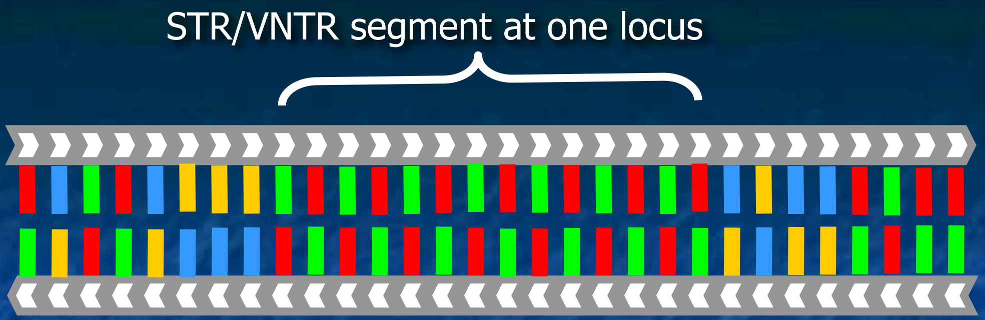
- Cool to ~ 65 C: **Primers** attach to single strands at locations encoded by their bases

Choose primer bases so they latch onto fixed DNA adjacent to target repeats!

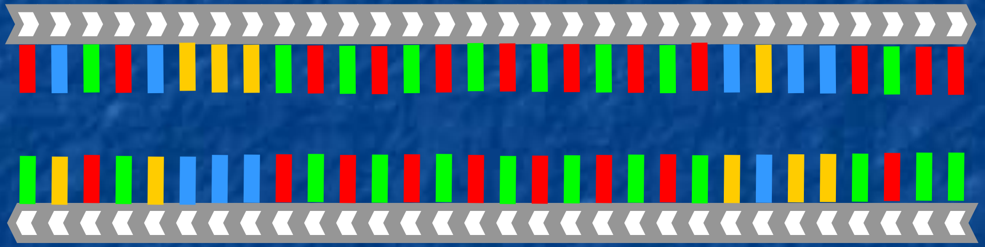
- Polymerase collects nucleotides, building complementary DNA from primer location

Easier to understand via figures:

1) Original complementary paired DNA strands

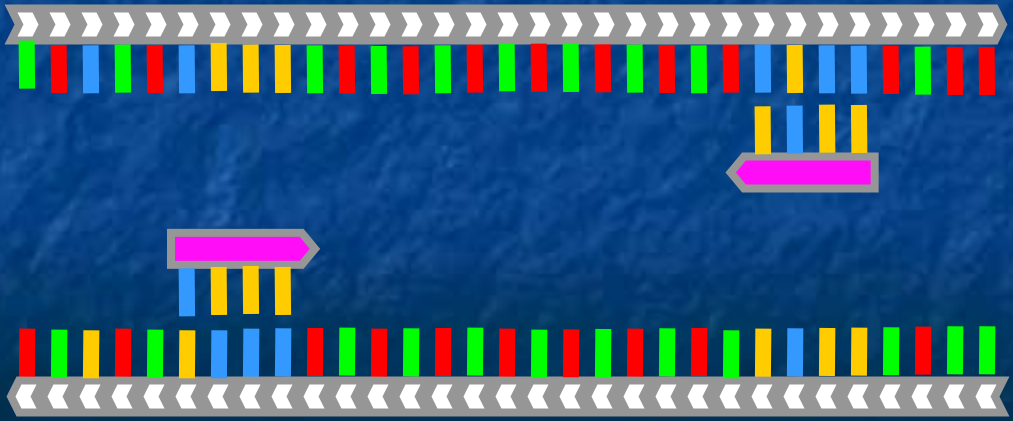


2) "Denature" (separate) pair by heating to 95 C



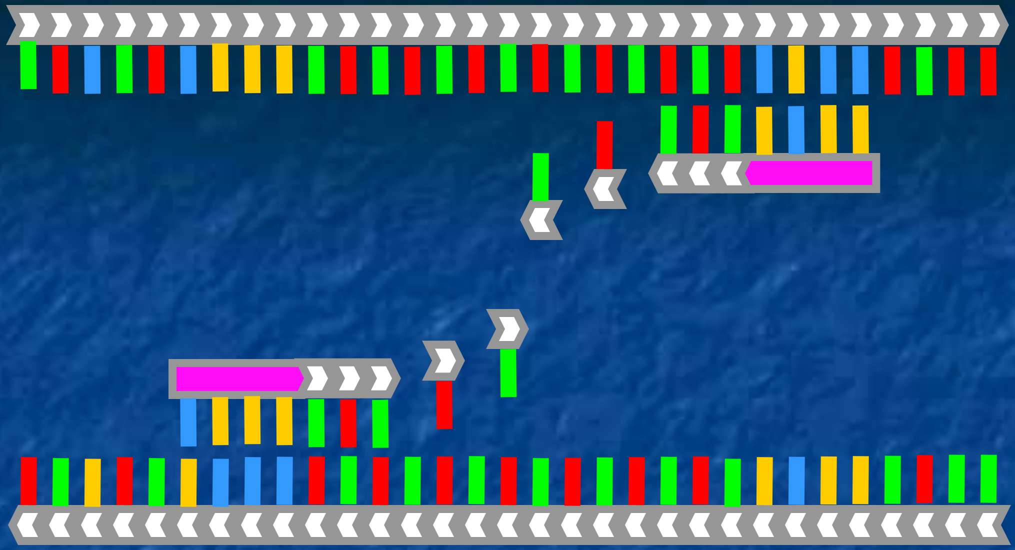
3) Cool to 65 C: Two "Primers" attach themselves upstream of the repeat sections (because of their base codes!)

It is the base coding of this primer pair that selects this particular repeat sequence for replication. To do this must have fixed known surrounding DNA.



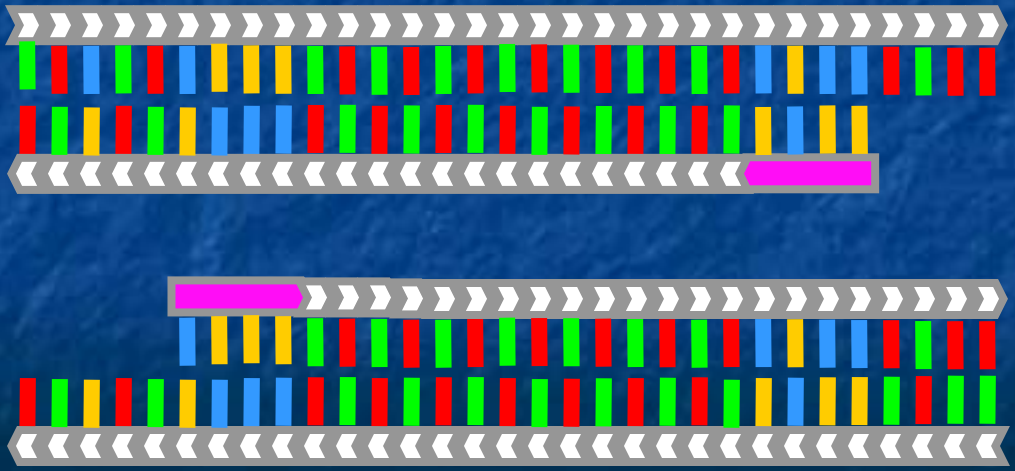
4) Complementary nucleotides drawn from solution to build from primers on the two separated single strands

(catalyzed by polymerase - not shown)



5) Completed 1st round of PCR

Note: Did NOT replicate ends upstream from primers

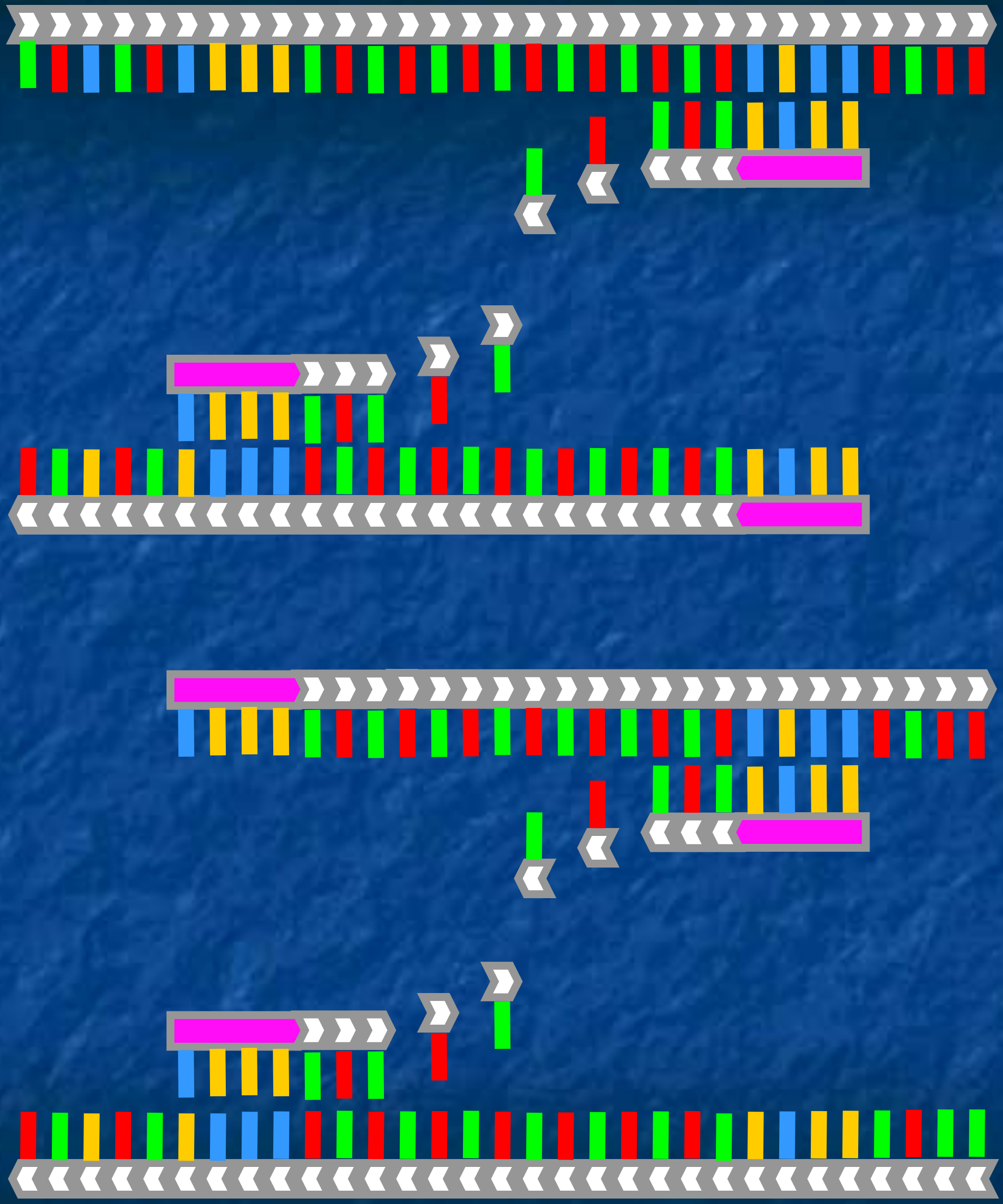


A Hands-on Introduction to Nanoscience: WeCanFigureThisOut.org/NANO/Nano_home.htm

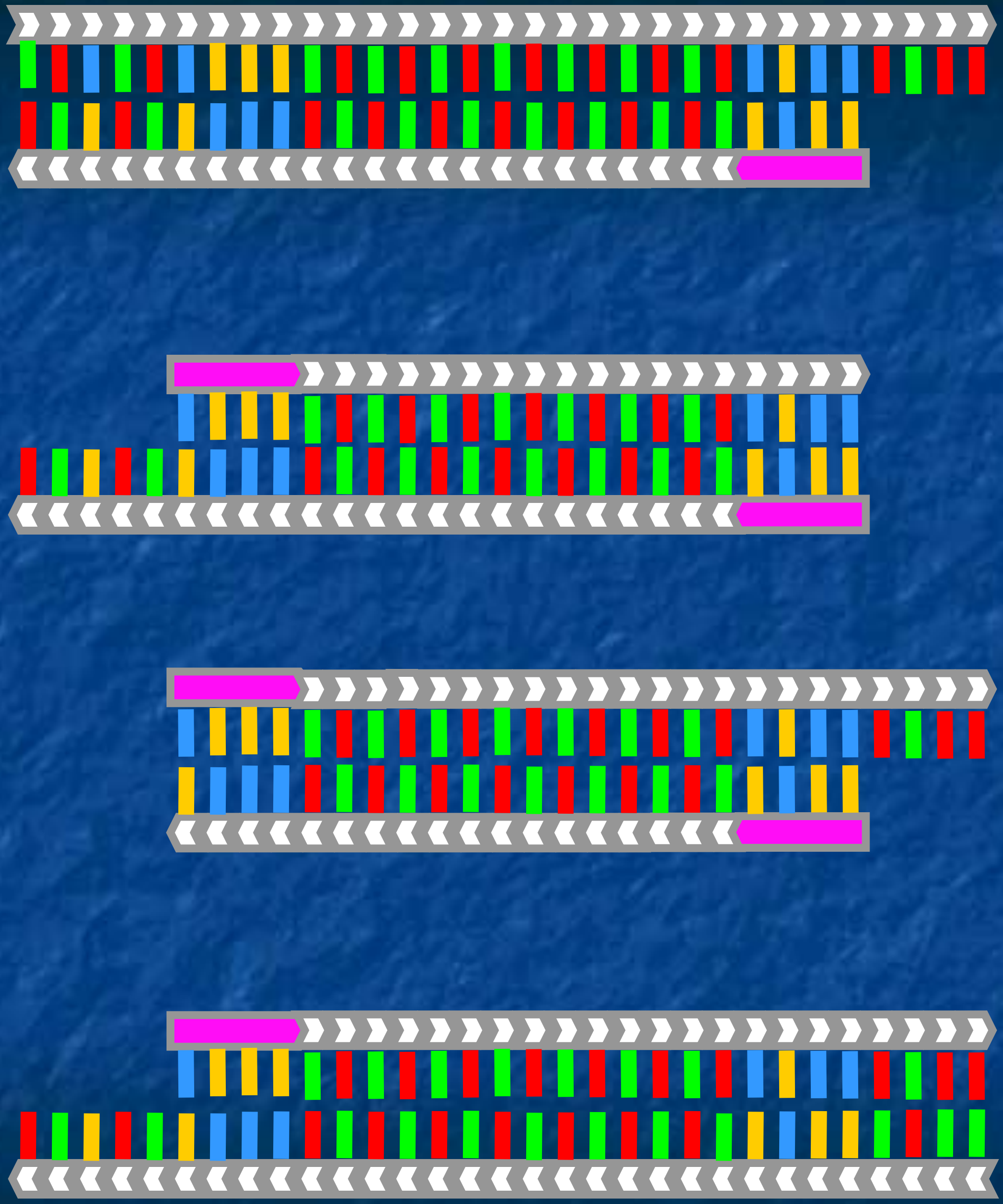
6 & 7) From two new pairs, replicate FOUR component strands:

- Denature at 95 C

- Prime & grow at 65 C



Yielding at the end of this 2nd PCR cycle:

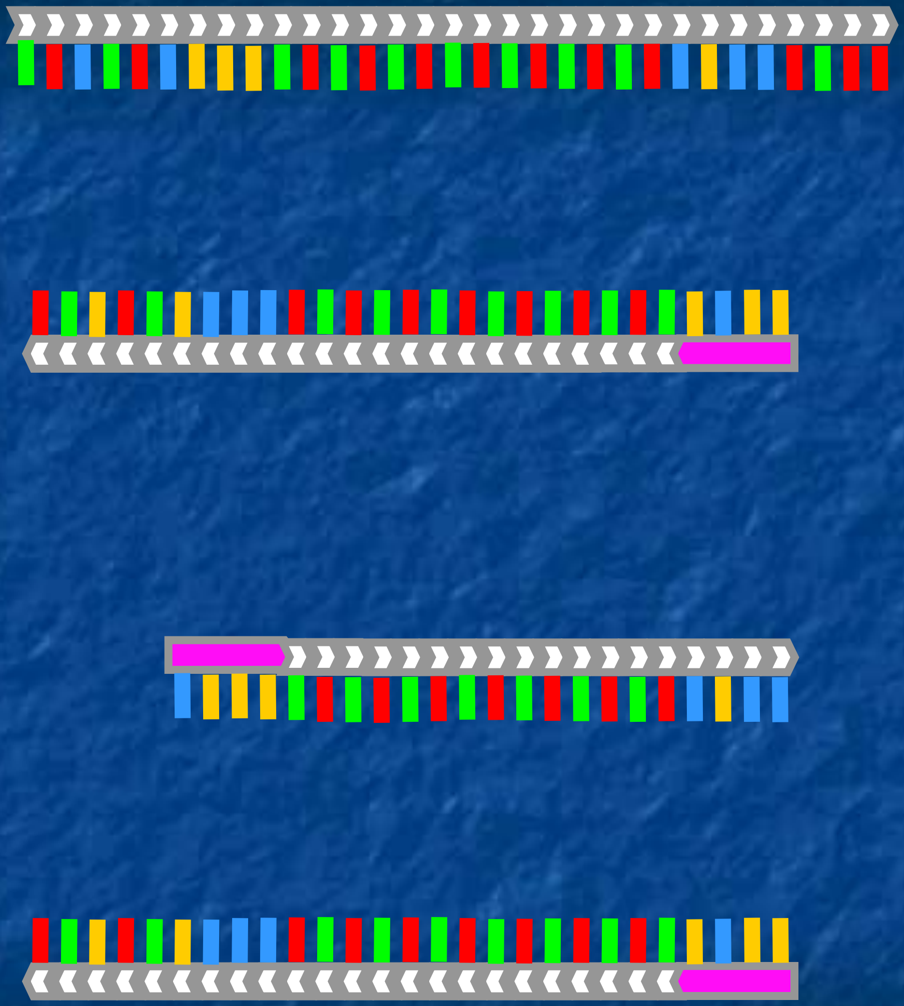


IMPORTANT TREND

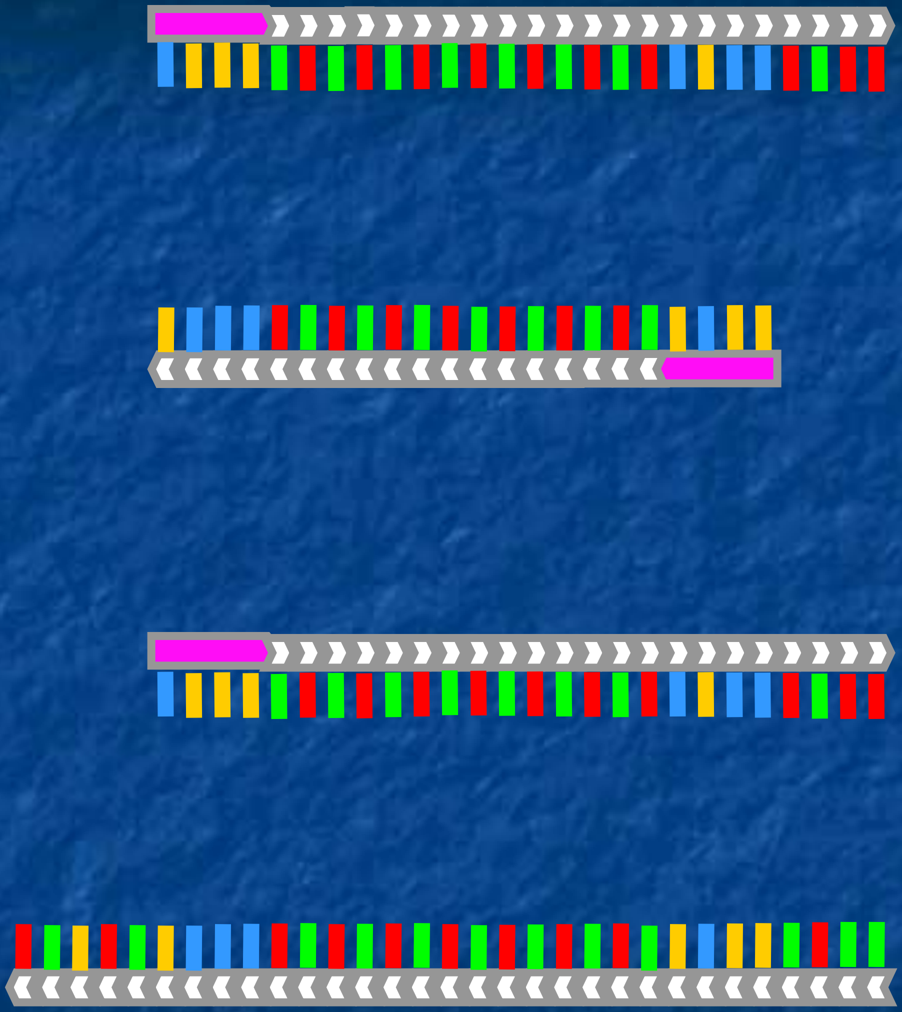
Because parts of strands upstream of the primers are not copied, the products are increasingly ONLY repeat sections + bases used by primers!

Here, in 3rd and 6th strands (25%)

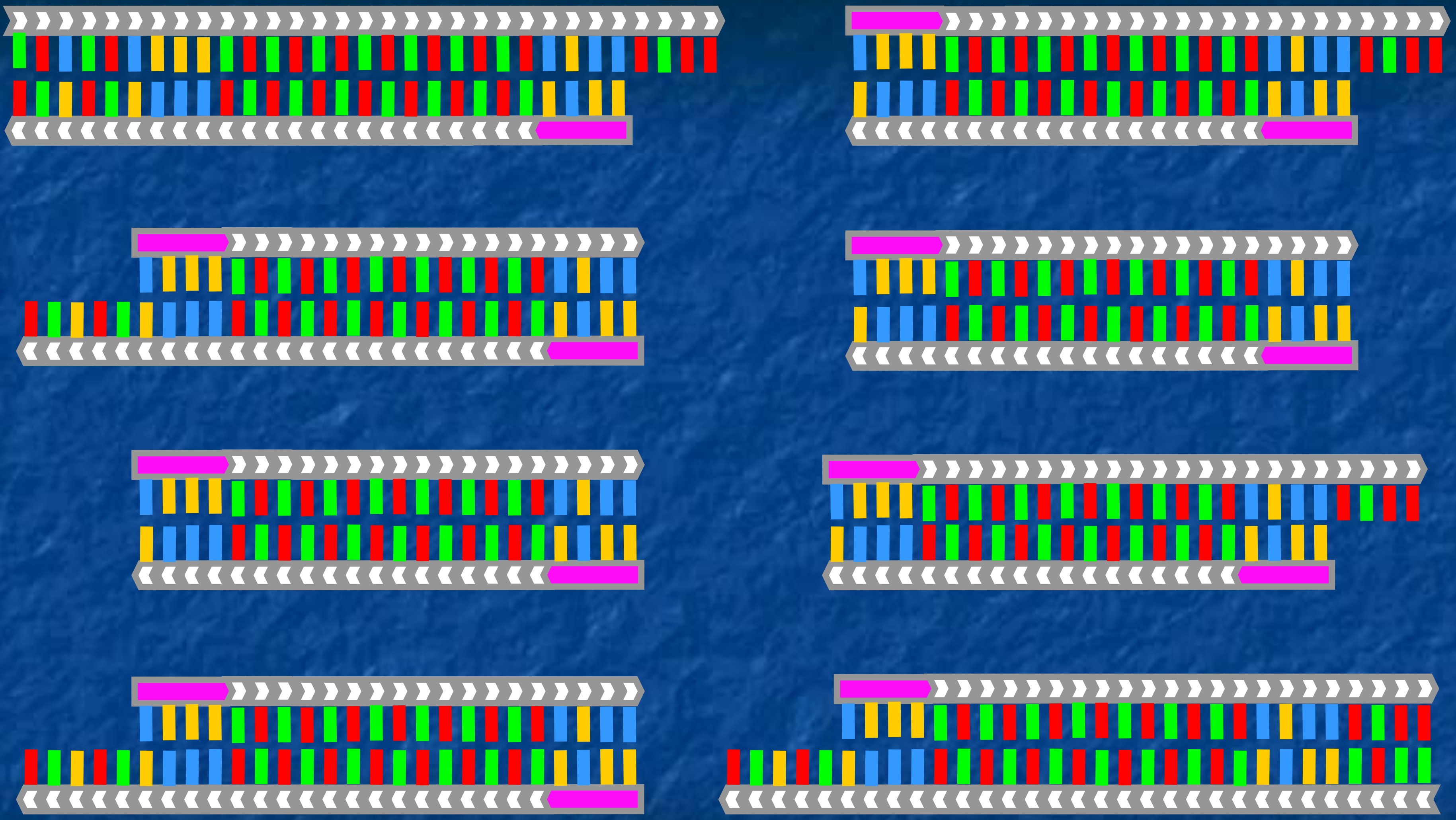
“Prove” trend with one more PCR cycle



Denature four pairs above to get:



After completion of 3rd PCR cycle:



Now 50% of strands consist of only repeat section + primer bases . . .

A Hands-on Introduction to Nanoscience: WeCanFigureThisOut.org/NANO/Nano_home.htm

Oversimplification Number One:

In my figures & preceding animation, primers are depicted as being VERY short

I showed them as including only four bases

But a four base code would be very non-specific, occurring all over the genome

Chance of match at random location ~ $(1/4)^4 = 0.39\%$

So to target ONLY a specific VNTR locus on the gene, primers MUST have more bases!

In our DNA fingerprinting lab, we will target a locus on the first chromosome called pMCT118

Our primers (intended to attach adjacent to the VNTR segment) have base sequences:

Primer 1:	5’-GAAACTGGCCTCCAAACACTGCCCGCCG-3’	Twenty eight bases!
Primer 2:	5’-GTCTTGTTGGAGATGCACGTGCCCCTTGC-3’	Twenty nine bases!

Chance of match at random location ~ $(1/4)^{28 \text{ or } 29} \sim 10^{-15}$

Oversimplification Number Two:

In my figures I also showed a VNTR repeating unit of only two bases

In fact repeat units (a.k.a. “consensus units”) can be much longer:

For our DNA fingerprinting lab targeting locus pMCT118, repeat unit is 16 bases long:

VNTR repeat unit: [G-(Any base)-(A or G)-G-A-C-C-A-C-(A or C)-G-G-(Any base)-A-A-G]

Along with a complementary segment on the coupled DNA strand

At this locus, most individuals have between 14 and 40 repetitions of this unit

Meaning that PCR at pMCT118 will amplify segments of base pair length:

(Primer 1) + (VNTR) x (14 to 40) + (Primer 2)

= 28 + (16) x (14 to 40) + 29

= 281 to 697 base pairs

Other research gives pMCT118 allele length range as 369 to 801 base pairs

But returning to the central points

PCR process accomplishes TWO things:

1) It selects out ONLY the target tandem repeat segment(s):

Primers designed to latch onto DNA to sides of repeat via their base coded addresses

Differently addressed primer pairs target different tandem repeat loci

So can PCR multiple tandem loci simultaneously

Just use different primer pairs for each locus

After many PCR cycles, product => ~100% tandem sequence + end primer bases

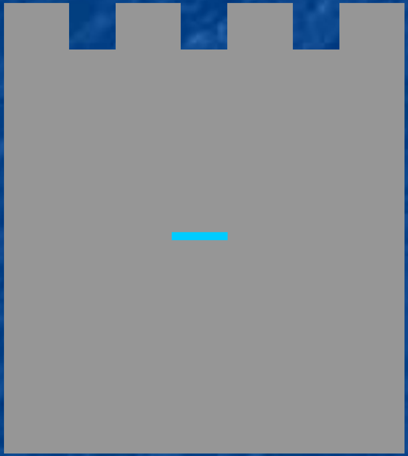
2) It made huge number of copies of the target repeat segment(s)

Replication process is called **amplification**

Then just apply electrophoresis to separate:

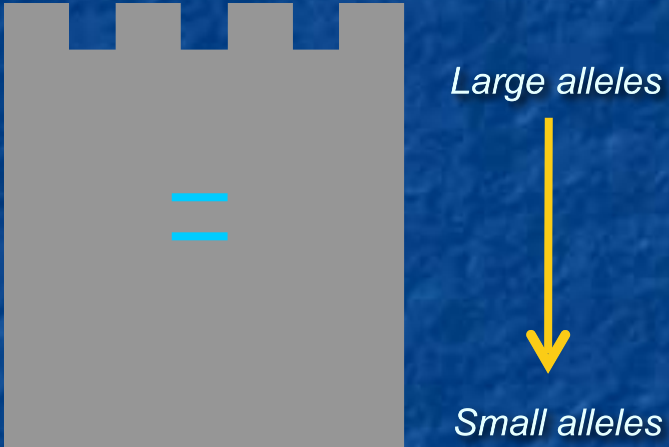
If used only single pair of primers targeting one tandem repeat locus:

Possible results after electrophoresis and blue staining of DNA:



With position of allele stripe(s) varying with allele size

of which there may be as many as 50 variations in the population as a whole



For individual with **homozygous alleles**

Same number of repeats on chromosome pair
(e.g. in 25% of population)

For individual with **heterozygous alleles**

Different number of repeats on chromosome pair
(e.g. in 75% of population)

Comparison between individuals?

Variations = Number of repeats at the targeted locus that occur in the whole population

For instance, might have: $(AG)^{21}$ $(AG)^{23}$ $(AG)^{24}$ $(AG)^{27}$ $(AG)^{30}$ and $(AG)^{32}$

But these variations might not be equally likely in population

Need some specifics:

For the pMCT118 VNTR PCR kit we will be using in the lab:

In whole population, there are 29 variations (alleles) of repeat number

(occurring with differing probabilities)

Net result is that there is a ~ 1 in 18 probability that two individuals have same alleles

=> 1/18 probability of matching PCR DNA fingerprint between individuals

1:18 chance of match? Does not compute!

That is because our lab kit analyzes only **one** tandem repeat locus

Via its use of a pair of PCR primers that are coded to latch on adjacent to this one locus

FBI's CODIS database instead analyses 13 different tandem repeat loci

That yields my earlier rough calculation assuming 10-50 variations at each of 13 loci:

Resulting total number of fingerprint variations ~ 10^{13} - 50^{13} = 10^{13} - 10^{22}

Approximate because:

- Number of variations (alleles) at each site (locus) may fall out of range 10-50

- Alleles for given site (locus) will not occur in population with equal probability

But is evident that analysis of 13 loci makes random match exceedingly unlikely!

So CODIS 13 locus PCR DNA fingerprints are infallible?

Depends on what you mean by infallible:

Numbers DO indicate spontaneous match between two individuals is exceedingly unlikely

But that does not rule out the possibility of a fabricated match:

"DNA Evidence CAN be Fabricated" - New York Times, August 18, 2009:

CODIS tests 13 loci: Using my range of 10-50 alleles per locus

=> 130 - 650 total number of alleles over 13 locus sites

Actual exact number, according to article, turns out to be 425

So to fake CODIS result:

Just acquire and maintain (via PCR) library with samples of each of these 425 alleles

If you then knew the results of a crime scene DNA analysis (= alleles of criminal at 13 loci)

You could then easily select the necessary alleles: (1-2 per locus) x 13 = 13 - 26 alleles

Then plant this synthetic DNA sample

(matching your “prime suspect”)

on objects that would implicate that person

I’m NOT suggesting any criminal investigator has ever done this

But there HAVE been documented cases of evidentiary fraud:

"FBI overstated forensic hair matches in nearly all trials before 2000"

Washington Post – 19 April 2015

"The Justice Department and FBI have formally acknowledged that nearly every examiner in an elite FBI forensic unit gave flawed testimony in almost all trials in which they offered evidence against criminal defendants over more than a two-decade period before 2000.

Of 28 examiners with the FBI Laboratory's microscopic hair comparison unit, 26 overstated forensic matches in ways that favored prosecutors in more than 95 percent of the 268 trials reviewed so far, according to the National Association of Criminal Defense Lawyers (NACDL) and the Innocence Project, which are assisting the government with the country's largest post-conviction review of questioned forensic evidence.

The cases include those of 32 defendants sentenced to death. Of those, 14 have been executed or died in prison, the groups said under an agreement with the government to release results after the review of the first 200 convictions."

Leading to final topic of weaknesses in PCR fingerprinting:

One of PCR fingerprinting's BIG advantages is that it "amplifies" small DNA samples

But this is also a potential weakness: PCR amplifies DNA at target loci from ANY source

Minute initial DNA contamination ALSO amplified => strong fingerprint lines!

Instead, minute DNA contamination of earlier RFLP technique => ~ invisible fingerprint lines

So not only must purity of PCR DNA samples be strictly "policed"

in practice, PCR analysis must also begin as soon as possible after isolation of DNA

These necessities led to much of the controversy over DNA analysis in 1990's

And to major investment/upgrades in DNA acquisition and testing procedures

Credits / Acknowledgements

Funding for this class was obtained from the National Science Foundation (under their Nanoscience Undergraduate Education program).

This set of notes was authored by John C. Bean who also created all figures not explicitly credited above.

Copyright John C. Bean

(However, permission is granted for use by individual instructors in non-profit academic institutions)