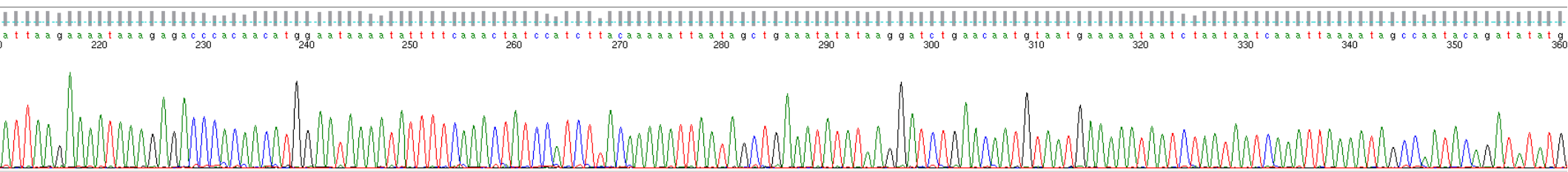


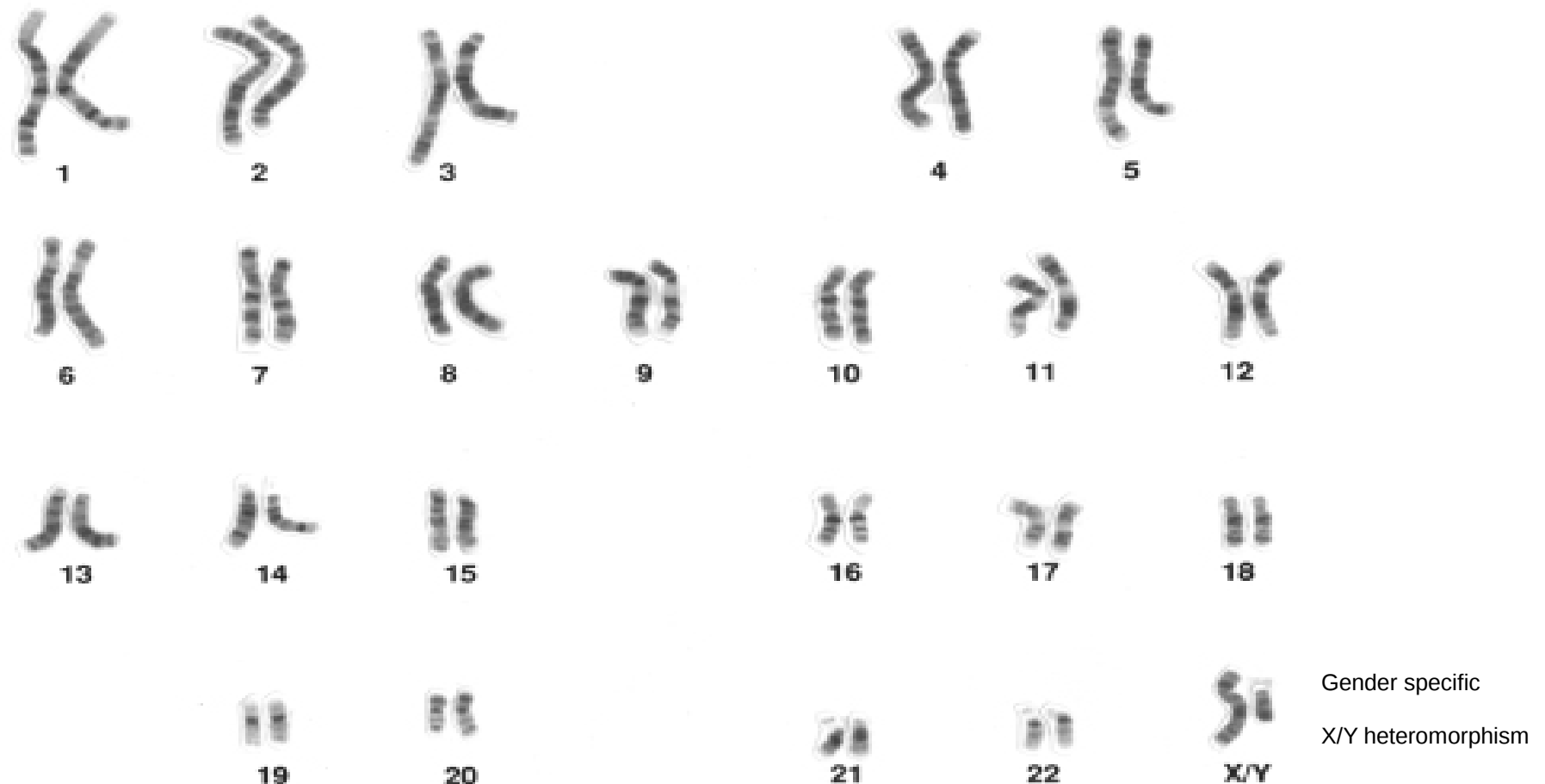
I've Received My Y Chromosome Sequencing Results – What Now?

During the past year in particular, new technologies have made it economically feasible to sequence large portions of the Y chromosome. This presentation will review the interpretation of Y chromosome sequencing data obtained by using various different technologies. Thomas Krahn will explain the different sequencing methods, as well as their advantages and disadvantages. He will describe how to combine them in an intelligent way to extract the correct and essential information from the overwhelming datasets. Thomas will also explore the implications of recombination events within different regions of the Y chromosome and with other chromosomes such as the X chromosome and demonstrate how this affects the interpretation of Next Generation sequencing data.



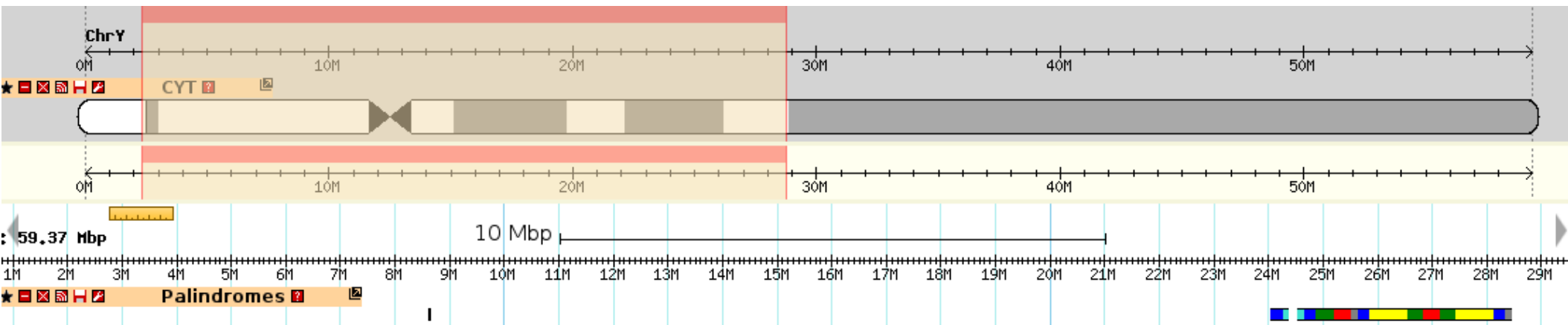
Y is an Unusual Chromosome

- Human Male Karyotype

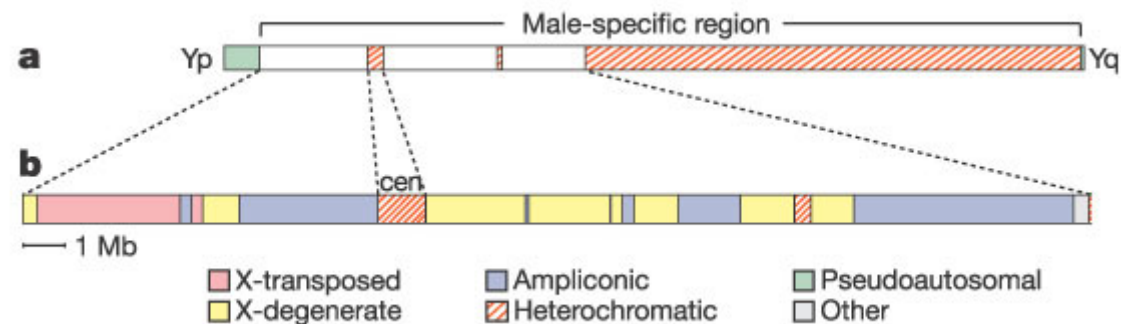


Courtesy: National Human Genome Research Institute

Y Has a Mosaic Of Discrete Sequence Classes



ISOGG Ybrowse (2014)

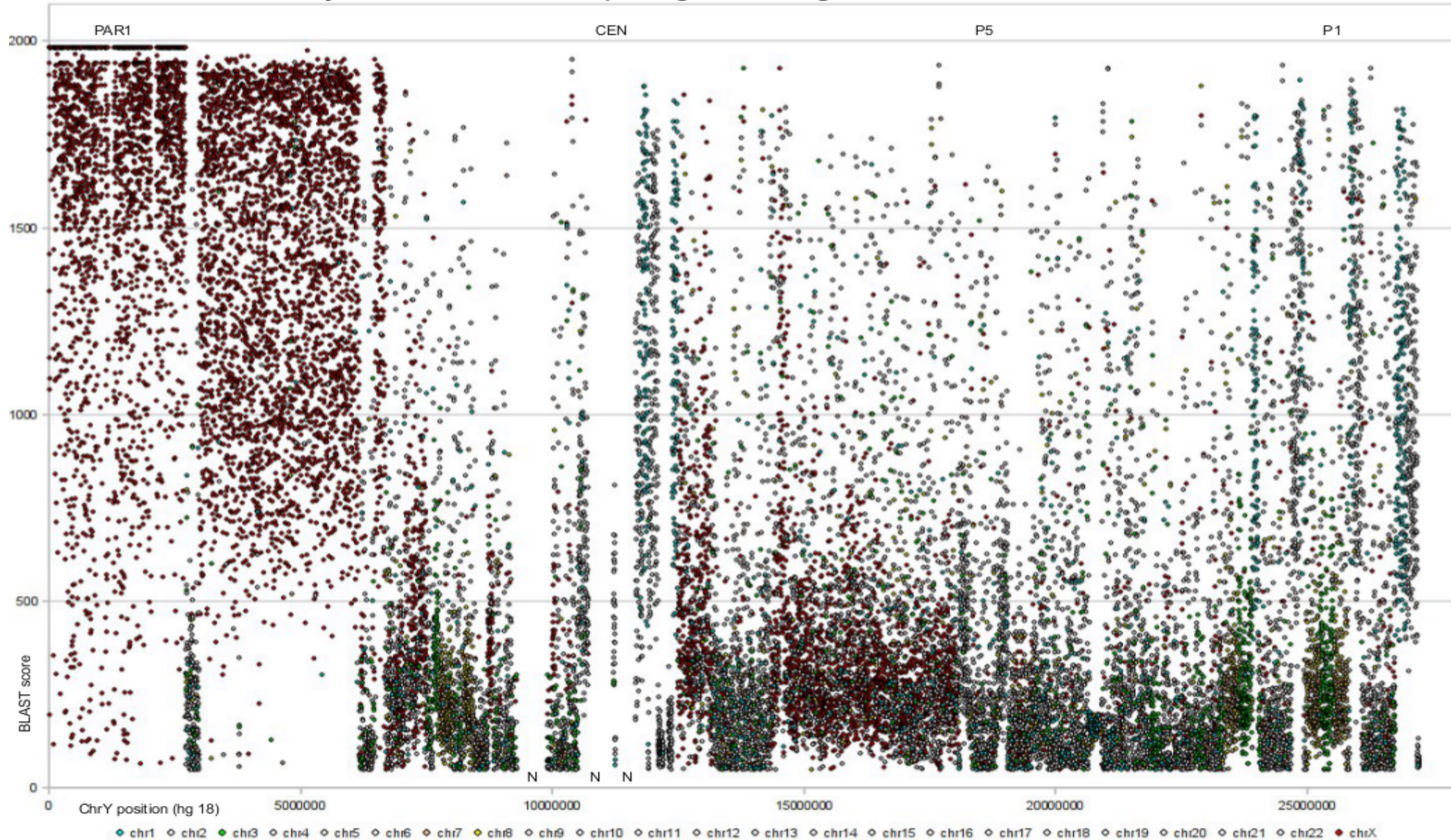


Skaletsky et al. (2003) Nature 423, 825-837

Y Has Highly Similar Regions to ChrX

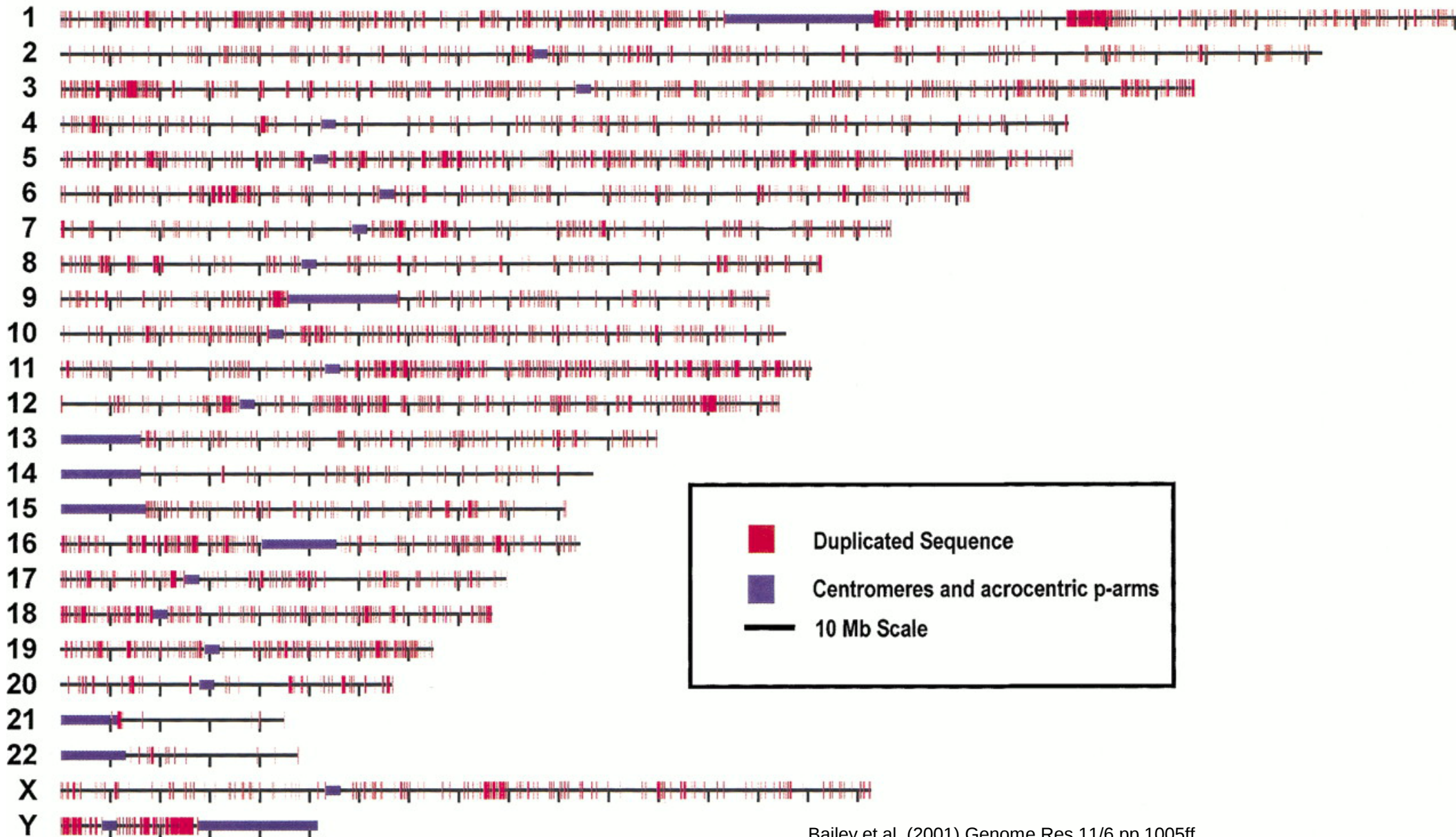
ChrY Similarity

Determined by BLAST of 1000 bp Segments Against All Other Human Chromosomes

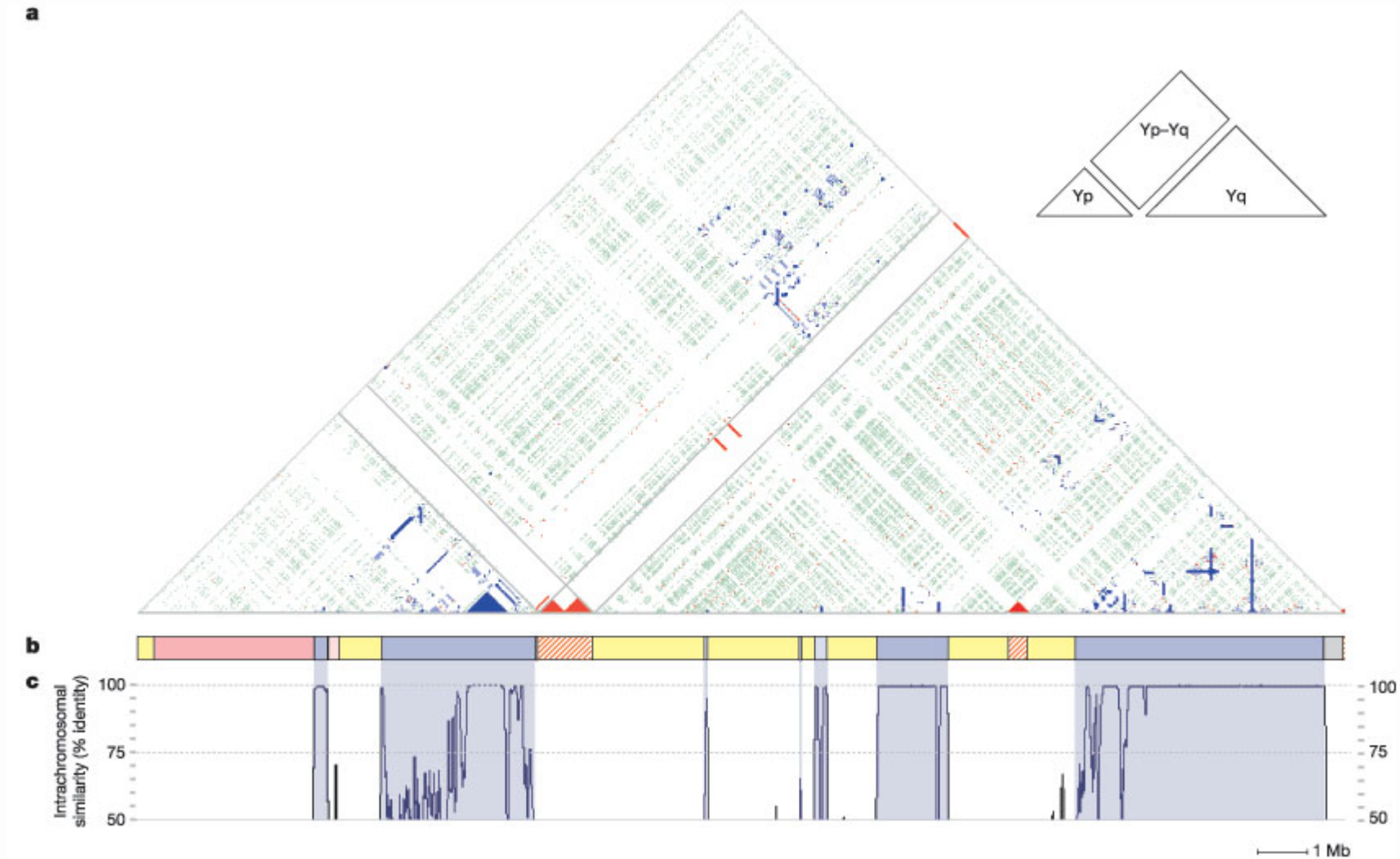


Y Duplications In Comparison to Other Human Chromosomes

B



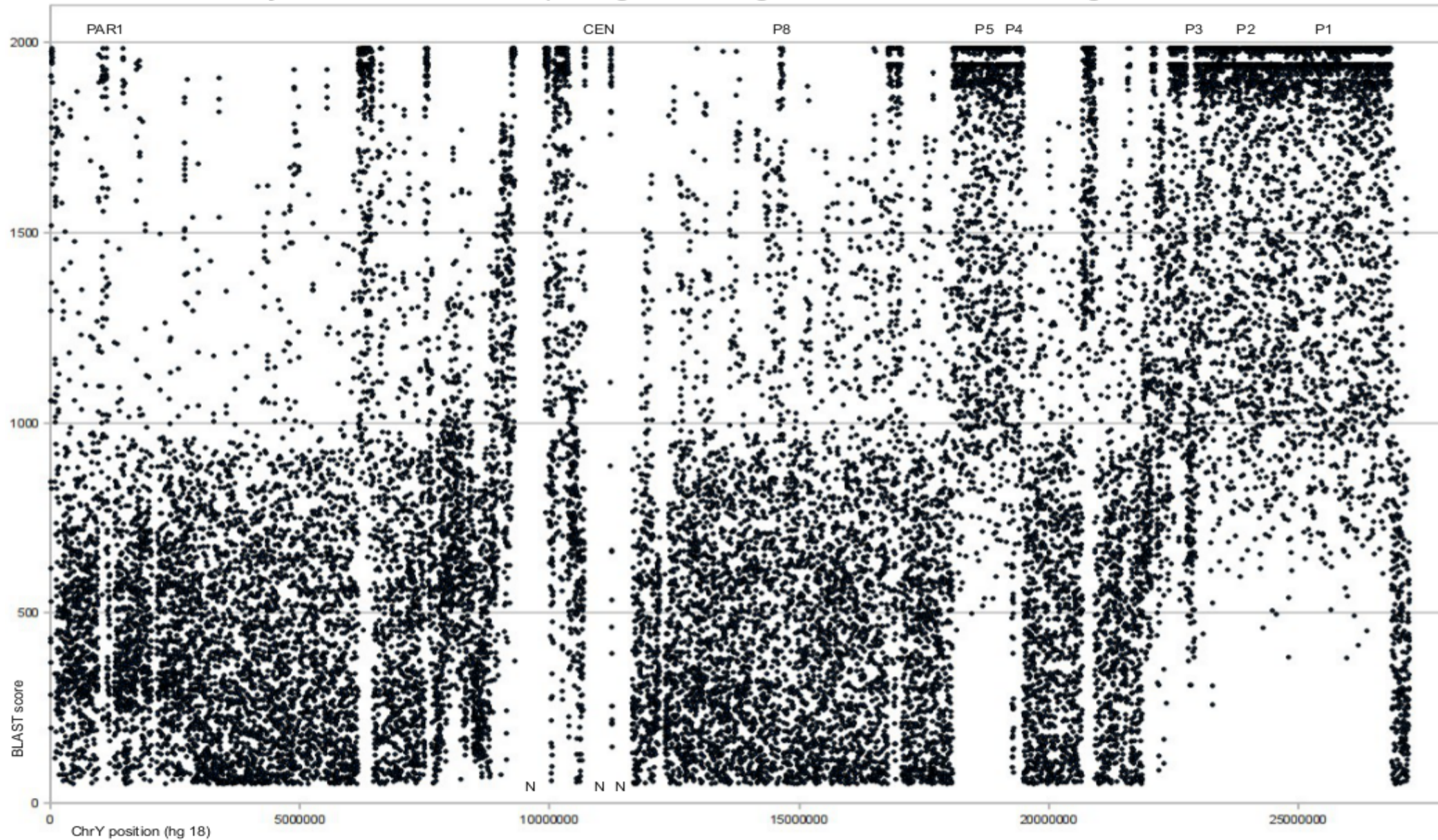
Y Has The Most Self-Repetitive Elements Of All Human Chromosomes



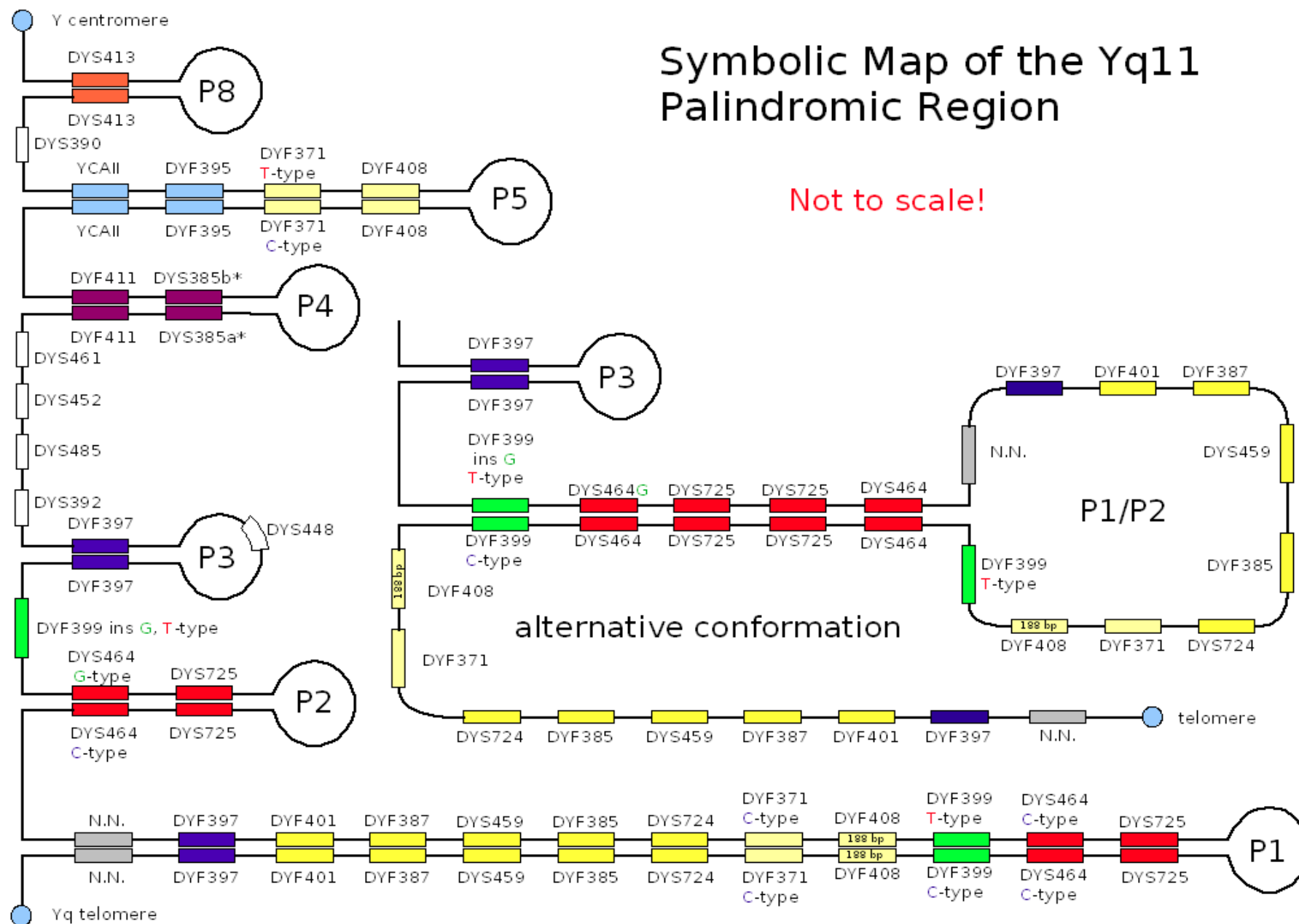
Y Self-Identical Regions

ChrY Self Similarity

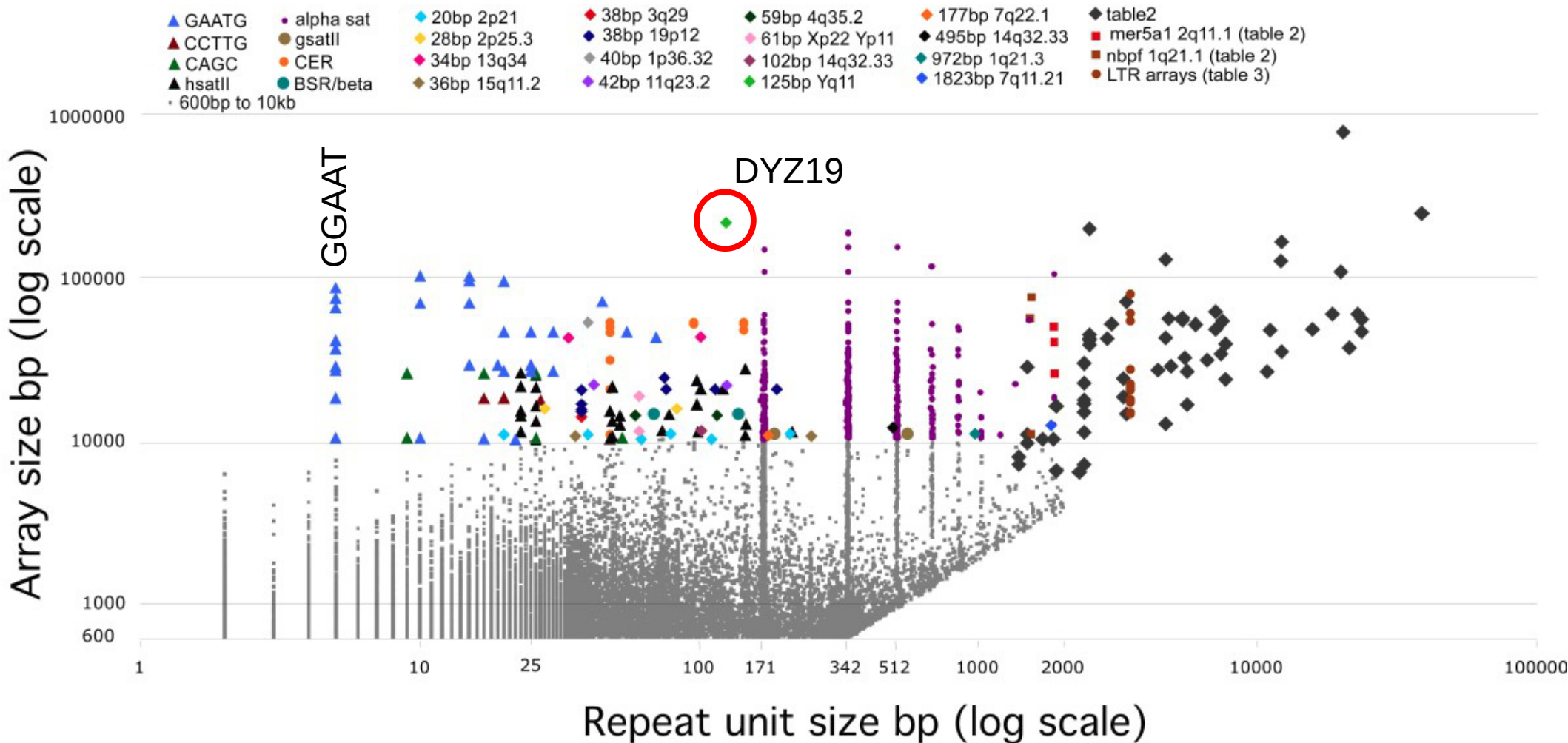
Determined by BLAST of 1000 bp Segments Against Itself And Picking The Second Best Hit



Yq11 Palindromic Region



Tandem Repeats Of the Human Genome



DYZ19 (125 bp) Repeat

A657

125 bp repeat can be seen with the plain eye already

```
>ChrY: 22260982..22261982
TGGTCTTGCTCACTTCAAGAATGAAGCTCCAGTCCTTTACGGTGAGTGTTACAGCACTTAAAGGTGTTATGTCCAGAGCTTTTTCTTCAGATGTGTCCAGAGTTTCTTCCTTCTGGCAGGTTCA
ATGGTCTTGCTCACTTCAAAAATGAAAGCTGCAGACGATTACGGTGAGTGTTACAGCACTTAAAGGTGTTATGTCCAGAGTTTCTTCAGATGTGTCCAGTTTCTTCCTCCTGGCAAGTTTCAT
GGTCTTGCTCACTTCAAGAATGAAAGCTGTAGACCTTTACGGTGAGTGTTACAGCACTGAAAGATGTTATGTCCAGAGTTTTCTTCAGATATTTCCAGAGTTTCTTCCTCTGGCAGGATCAT
GGTCTTGCTCACTTCAAGAATGAAGCTGCAGACCTTAGTGGTGAGGTTACAGCACTTAAAGGTGTTATGTCCAGAGTTTTCTACAAATGTGTCCAGAGTTTCTTCCTTCTGGCTGGTTTCAT
GGTCTTGCTCACTTCAAGAATGAAGCTACAGACCTTAGTGGTGAGTGTTACAGCACTTAAAGGTATTATGCCAGGGTTTGTACTTCAGATGTGTCCAGAGTTTCTTCCTTCTGGCAGGTTCA
TGGTCTTGCTCACTTTAAGAATGAAGCTGCAGACATTTACGGTGAGTGTTACAGCACTTATATTTGTTGTATCCAGAGTTTGTTCCTTCAGATGTGTCCAGAAATTTCTTCCTTCTGGCGGGTTCA
TGGTCTTGCTCACTTCAAGAATGAAGCTGCAGACCTTAGGGTGAGTGTTACAGCACTTAAAGTTTTATGTCCAGAGTTTGTTCCTTATAGATGTGTCCAGAGTTTCTTCCTTCAGGCAGGTTCA
TGGTCTTGCTCACTTTAAGAATGAAGCTGCAGACCTAGGTGGTGATCGTTACAGCATTTAAAGGTGTTATGTCCAGAGTTTTCTTCAGATGTGTCCAGAGTTTCTTCCTTCTGGCAGGTTTCAT
G
```

Search

Landmark or Region:

ChrY:22,216,800..22,512,940

Search

Download Decorated FASTA File

Configure...

Go

Save Snapshot

Load Snapshot

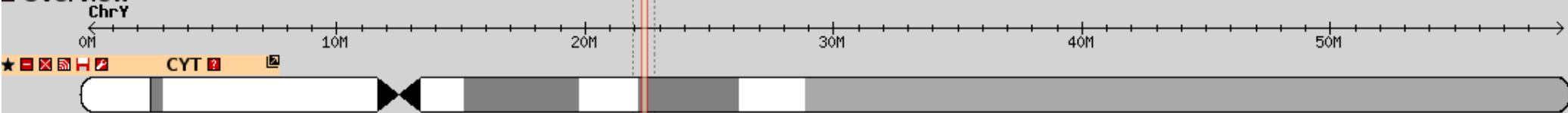
Examples: ChrY, ChrY:15000000..15000100, L21, DYS437.

Data Source

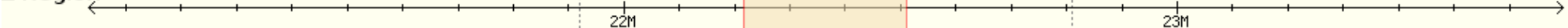
Human Y Chromosome

Scroll/Zoom: << < - Show 296.1 kbp + > >> Flip

Overview



Region

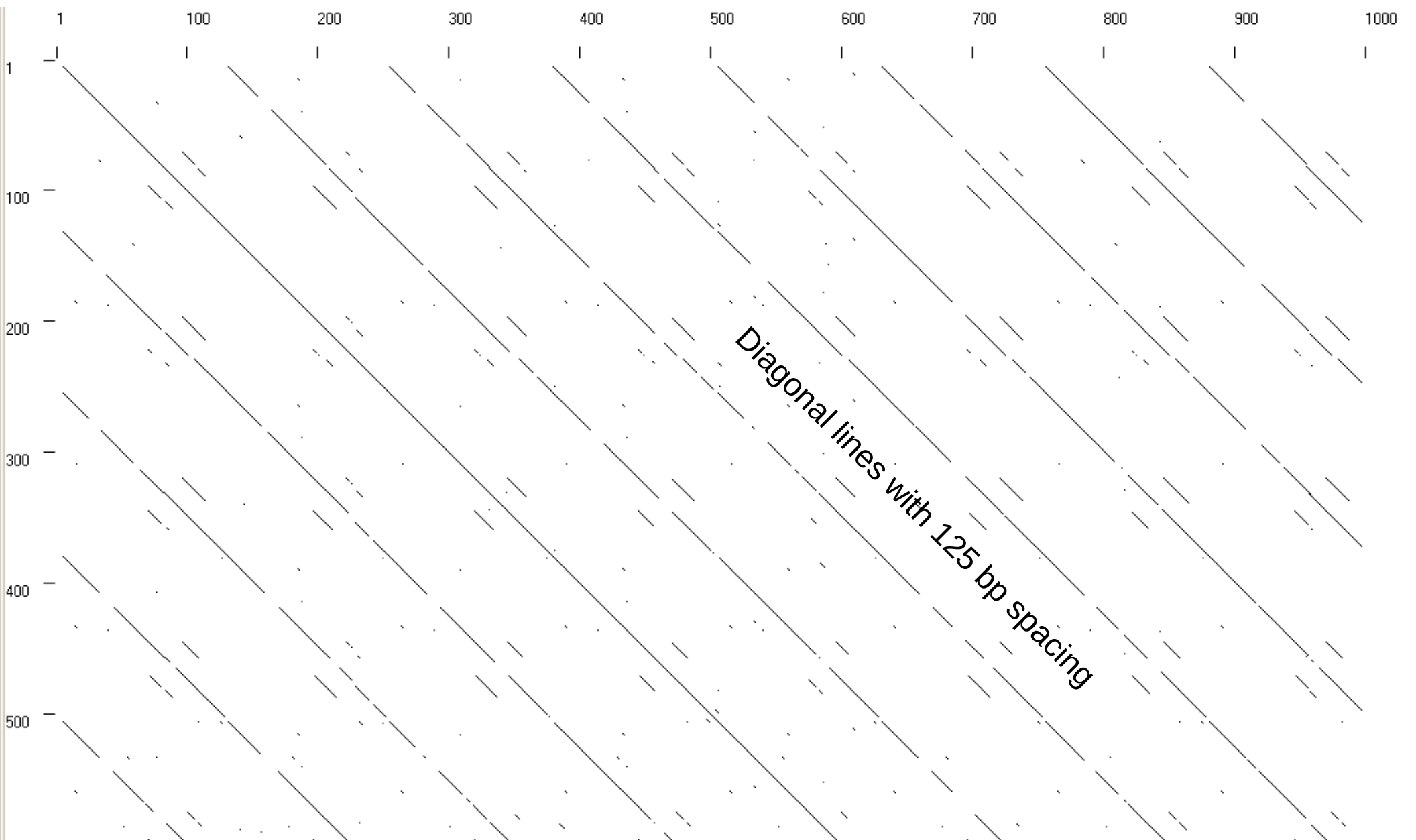


Details



DYZ19 Dot Plot

Segment around A657 (+/- 500 bp, ChrY:22260982..22261982)

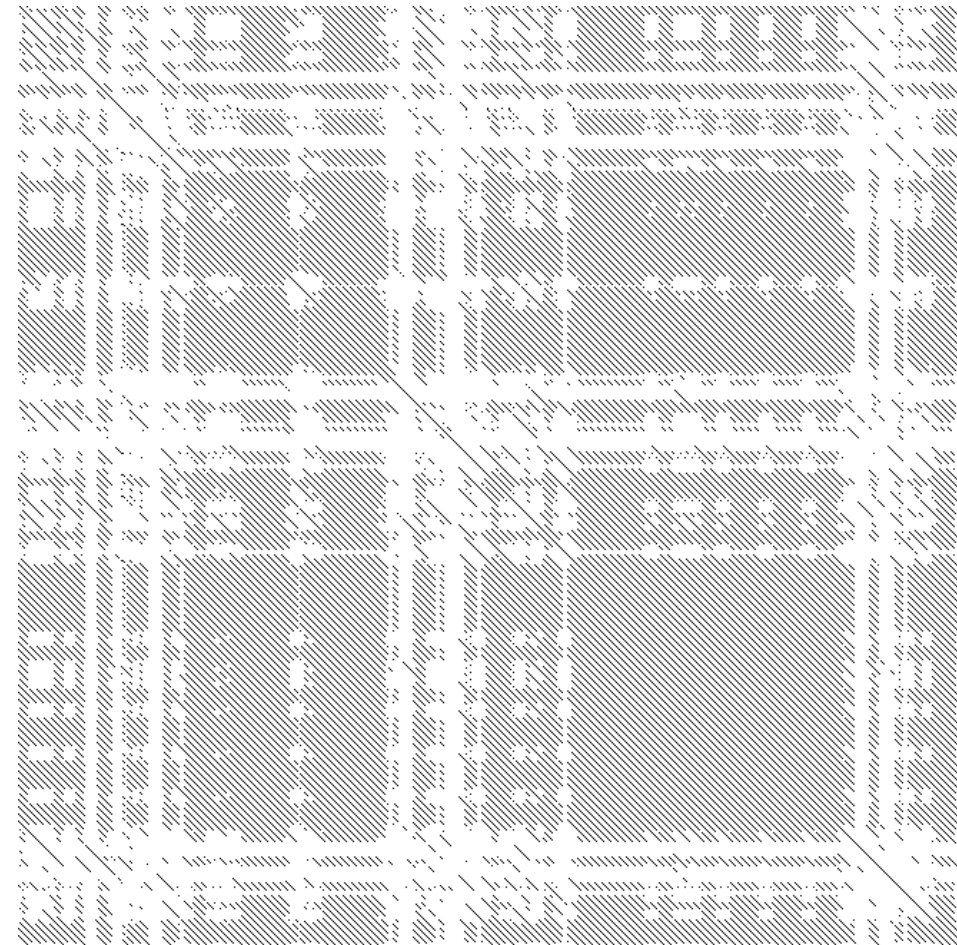


Ycen and Yq12 GGAAT Repeats

>ChrY:13810160..13812159

aatggcctcgaagaatggataggaatggaatggactcgaatggattggactcgaatgg
aatagattagaattgaatgacatcaaatggaatggaatggaaccgaatgtaatggactcg
aatggaatggattcaaatgaaaggactcaaaagggaatgggtcccgaagggaatttattcg
aatagaatggaatcaaacagaatgcagtagtagggaatggaattgaattcaaaaggatcgc
aaaagaatggaccggaatggaacagtcgaagtatagaacagaatcgaatgtaatggattgc
aatgcaattgggttcaaatggaacggatcgaatggaatgtaatcaaatggaatggaagg
aattgaatggattacaatggaacggaacggagtggaatcgagggtaatggaatctaattt
aatggaatcgaatagaatggactggaatggaatggacacgaaagggaatggactggaagaa
aatggaattgaacggattggaatcgagcgggaatggaaggtaatggaatggattagaatgg
agtcgaatggaatggagttgaatggaatggaccggaatggaatggaatcaaaaggaaatgg
actggagtggaatagacacgaaagggaatggaaacgaatgcaatggaatggaatggaagg
aatgcaatggaaagggaatggaatggaatggaatggaatggaatggaatggaatggaattg
aaagggaatgcaatggaatggaatggaatggaatggaatggaatggaatggaatggaattg
aattggatggaacaaatggaattgaatggagtcgaatagaatatcaaaagggaatgg
catcgaagggaatggaatggaaagggaatggattcgaatggaatgcaatggaatggaatgg
tattgaaagggaatggaatggaatggaatggaatggaatggaatggaatggaatggaatgg
aatgcaatggaatggaatggaatggaatggaatgcaatggaatggaatggaatgcaatgg
aatggaatggaatggaatggaatgcaatggaatggaatggaatggaatggaatgcaatgg
aatggaatggaatggaatggaatagaatgtactcgaatggaatggactcaaaagggaatag
aatcgtatggaatggcatcgaatggaatggaatggaatggaccacataattgactcg
aaaagaatagacacaaatagaatggacttgaaagggaatgggtttcgaatggaatttattcaa
atagaattgaatcaaatggaatgtaatagtatggaatgaaatcaaatggaatggaatcga
aatgaatggaccggaatgcattggactggaatgaaagggaatcgaatggaatgggttacaa
atgtaattgattcaaatggaatggaatcgaatagaatggaatcaaatgggattggaatgga
atgcaatgtaatggaaggtaatgcaatgtaatggaacagaatggcatacaatggaatgga
atggaatggagtggaatcgagtggaatggagtcgaatggaatgggtattgtgtggaatgga
ttggaattgaatggactcgaatggaatggaatggaaaaaatggaatcaaatggatagga
atcgaacggaacggaatagaatggaatggaatcgaatggaatggagtcaaaagggaatgga
atcgaatggaatggaatcgaatgtaatggaatatcaatggaatgaaattgaaagggaatcga
aagggaatagaagtgaatggagtgtaatggaatgacattgaatggaatggattggaatgga
ctcgaatggaacggaatggagtggaatcgacttggaatggaatggactggagtggaatgga
ttcgagtggaatggaacaaatggaatggaatgcaatggaaagggaatggaatggaataga
attaaatggatttggatggaacggaatgcaatggaatggagtcgaatggaatagaatcga
aagggaatggcatcaaatgca

Dot-Plot of the sequence (left)



What's Wrong With Highly Identical Regions?

- Fewer differences => Higher probability of false mapping of short reads
- More difficult to design specific PCR primers because they need at least 1-2 bases difference at their 3' end
- Highly similar regions are prone to recombination events
 - Parallel and back mutations
 - RecLOH at palindromic arms and tandem repeats
 - X-Y recombination (or other chromosomes)
 - Duplications, deletions, inversions
 - Large scale chromosomal rearrangements
 - STR like frame shift mutations
- => Highly identical regions screw up the phylogeny

So Should I Throw Away All This Highly Identical Regions?

- Identical regions are not necessarily bad data (unless mapping errors are made)
- However they need a lot more caution and background knowledge when interpreted within a phylogenetic context
- Those regions are expensive and time consuming to analyze and it may turn out that they deliver misleading information because recombination events were not considered
- => Prioritize your research and focus on stable regions first, where it is unlikely to cope with recombination events
- But of course keep the data in a safe place so that you can address the more difficult data later when new analysis techniques are developed and it becomes cheaper to analyze them

Where Do I Find The Most Reliable SNPs?

- Not in palindromic arms
- Not in the centromere
- Not in Yq12
- Not in the pseudoautosomal regions
- Not in the highly X-Y homologue region (large part of Yp11.2)
- Not in DYZ19 (125 bp repeat)
- Not inside short repetitive elements (e.g. STRs or homopolymers)
- Not flanking short repetitive elements
- Not in the “post-palindromic” region

What's Wrong with The “Gold Standard” Regions?

- Most numbers end with “000” => Low resolution
- Misses important known phylogenetic markers (e.g. L2: ChrY:5755550, DF97: ChrY:21040762)
- Doesn't consider short repetitive elements like STRs or homopolymers
- Assumes that all Y chromosomes have the same large scale arrangements like the reference sequence (no inversions, duplications or deletions allowed)
- Includes the “post-palindromic” region
- Article is behind a paywall
- Neat simplification of a complex problem, but by far not a “gold standard”

chrY	2655000	2674000
chrY	2680000	2913000
chrY	6619000	6721999
chrY	6726000	7016000
chrY	7034000	7114000
chrY	7127000	7140000
chrY	7144000	7431000
chrY	7519000	8003000
chrY	8007999	8905000
chrY	8976000	9156000
chrY	9393000	9428000
chrY	9431000	9461000
chrY	9798000	9895000
chrY	13885000	13976000
chrY	13984000	14430000
chrY	14447000	14960000
chrY	14965000	15213000
chrY	15253000	15284000
chrY	15303000	15636000
chrY	15643000	15666000
chrY	15682000	16011000
chrY	16015000	16067000
chrY	16187000	17973000
chrY	18031000	18256000
chrY	18554000	19163999
chrY	19188000	19381999
chrY	19391000	19483000
chrY	19493000	19551000
chrY	21049000	21150000
chrY	21158000	21744999
chrY	21746000	21828000
chrY	21843000	21942000
chrY	21944000	22209999
chrY	22541000	22628000
chrY	22645000	23156000
chrY	23242000	23637000
chrY	23746000	23898000
chrY	23958000	23991999
chrY	24373000	24386000
chrY	24388000	24480999
chrY	28477000	28551000
chrY	28598000	28770999

Poznik et al. (2013), science 341 (6145): 562-565

How Can I Properly Grade My SNP Candidates?

- Usually you'll find just a limited number of good quality SNP candidates (<200), so simply verify them one by one
- Select a region around the SNP candidate (e.g. +/- 500 bases) and BLAST/BLAT it to the complete reference genome
- If you find hits >95% identity (other than the original sequence itself) put this in the category “not resolvable today” and store it for a later time.
- If you find hits >85% identity, then add them to a multi sequence alignment and check where they are different.
- Examine the proximity of the SNP candidate for short repetitive elements. >8 repeats of any kind become dangerous!
- If you are a software engineer write a script that does this all automatically



Example A353

BLAT

Zoom out further in Ybrowse to 1kbp and download “Decorated FASTA File”
 Paste this in UCSC BLAT (<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>)

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	ChrY:19216074..19217073	1000	1	1000	1000	100.0%	Y	+	19216074	19217073	1000
browser details	ChrY:19216074..19217073	294	576	992	1000	87.5%	X	+	4200071	4200469	399
browser details	ChrY:19216074..19217073	281	57	637	1000	88.8%	10	-	9724443	9962175	237733
browser details	ChrY:19216074..19217073	266	57	643	1000	82.7%	8	+	91860768	91861295	528
browser details	ChrY:19216074..19217073	234	58	646	1000	90.2%	6	-	164279960	164696992	417033
browser details	ChrY:19216074..19217073	213	57	703	1000	91.8%	9	+	10898455	11212869	314415
browser details	ChrY:19216074..19217073	211	57	707	1000	82.6%	5	-	112307833	112308146	314
browser details	ChrY:19216074..19217073	204	12	638	1000	82.0%	X	+	42824316	42824605	290
browser details	ChrY:19216074..19217073	203	25	647	1000	88.1%	12	-	120382419	120383035	617
browser details	ChrY:19216074..19217073	201	36	643	1000	83.1%	6	-	11847366	11847637	272
browser details	ChrY:19216074..19217073	194	36	647	1000	82.3%	X	+	140402090	140402365	276
browser details	ChrY:19216074..19217073	193	57	645	1000	84.3%	14	-	86558854	86559290	437
browser details	ChrY:19216074..19217073	193	36	643	1000	80.7%	5	+	77740071	77740342	272
browser details	ChrY:19216074..19217073	190	12	643	1000	81.4%	8	+	54772197	54772493	297
browser details	ChrY:19216074..19217073	187	36	638	1000	79.8%	4	-	133352666	133352941	276
browser details	ChrY:19216074..19217073	186	59	500	1000	89.2%	9	-	15368102	15368679	578
browser details	ChrY:19216074..19217073	186	60	655	1000	82.6%	5	-	157296317	157296877	561
browser details	ChrY:19216074..19217073	184	36	638	1000	80.1%	4	+	178752811	178753074	26

No secondary match above 95% identity => good!
 However some matches with other chromosomes above 85%.
 So let's have a closer look at them.

Alignment of ChrY:19216074..19217073 and ChrX:4200071-4200469

cDNA ChrY:19216074..19217073

```

aggggtataa atattagttt attgtcatat cgctgtgaag aaatgcccaa 50
gcctgagtaa tttataaaga aaaaagggtt aatggactca cagttccaca 100
ttgtcgagga ggccctacaa tcatggcaga aggcaaagga ggagcaaagg 150
catgtcctac atggcaacag gcaagacagt ttctgcaggg aactctcctt 200
tacaaaaatga tcaagtcggc cagatgtggt ggttcacacc tctaatacca 250
gcactttttg aggtttgggt gggttgatca agatgtcagg agtttaagac 300
cagactggcc aacatactga aaccccatat ctactaaaaa tacaaaaaat 350
taccagaat gatggtgggc acttgaatt ctggctactc gagaggctga 400
gaaaggagaa tttcttgaa cttggaaggc gaagttgcag tgagctgaga 450
tcgcaccatt gcactccagc ccgggtgaca agagtgaaac tccatctcaa 500
aaaacaaaac aaaacaaaac aaaacaaaac aaaacaaaac aaaagccatc 550
aggctctgtg aaacttattc actgtCATGA GAACAGCATG GGAAAACgCG 600
CCCCAatgat ttcattacct cccaccAGGT CCCTCCCAgg acacatgagg 650
attgtgggaC TACAGTTcAa aATGAGATTT GGGTGGGAAC AtgGCCAAAC 700
CATGTCAGTA CAAGTtCAaC TTTTTTACAT GGAcATATTa tgtaGGGGTA 750
AAtTTGGGgC TTTtAGTGAA TTCATCACCT GAATAGTgTA CtTTGTACCC 800
AAaAGgTgAt TTCTCATCCC TCAaCCCTCT CCACaCTCCT AtCtTTCCAA 850
GTCTCCAATG TCTACTATTC CACTCTCCaT GTCCAATATGG ATAtATTATT 900
TAGCTCCAC TTATaaGGGa GAACAgGCAa TATTgGaAtt TGACTTTCTG 950
TTTCTGAGTT ATTTTCCTTA AGATAATGGt CTCCAGTTCC Atttgtgttc 1000

```

Only half of the segment is affected at all => good

The differences behind the STR allow for good discrimination of short NGS reads and there are enough difference “islands” where PCR primers can be designed for Sanger confirmation.

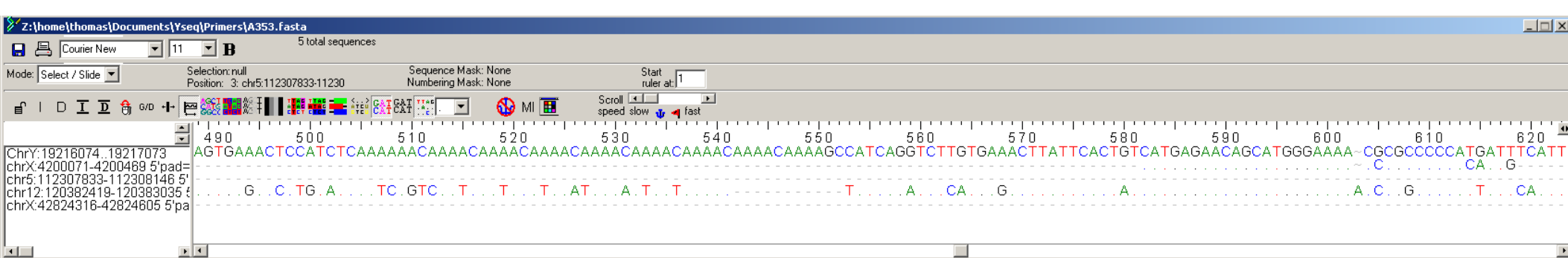
Genomic chrX :

```

aaagcaaagg aggaacatag gcatggcggc aggcaagaga gcttgtgcag 4200020
gggaactccc ctttataaaa ccatcaggtc tagtgagact tactcattaa 4200070
CATGAGAACA GCATGGGAAA ACcCGCCCCC AcaatgAGGT CCCTCCCAtg 4200120
actcatgggg attattggag CTACAGTTCA TGAGATTTGG GTGGGgACAg 4200170
aGCCAAACCA TGTCAGTACA AGTgCAgCTT TTTTACATGG AtATATTgca 4200220
tGGGGTAAAg TTGGGAcTTT gAGTGAATTC ATCACCTGAA TAGTtTAcCt 4200270
TGTACCCAAt AGcTaAcTTC TCATCCCTCA CCCTtCTCCA CcCTCCcAgC 4200320
cTTCCAAGTC TCCAATGTCT ACTATTCCAC TCTCCgTGTC CAcATGGATA 4200370
cATTATTTAG CTCCCACCTTA TgtGGGtGAA CAcGCAGtAT TaGtATaTGA 4200420
CTTTCTGTTT CTGAGTTATT TTCCTTAAGA TAATGGcCTC CAGTTCCATc 4200470
catgttgctg caaaagacat gattttcattc ttttttatgg ttgcatagta 4200520
ttccatggtg tacatgtacc acagtttctt catccagtca tccactgat

```


A353 Multiple Sequence Alignment with BioEdit



BioEdit is a great tool for investigating small DNA regions
<http://www.mbio.ncsu.edu/bioedit/bioedit.html>

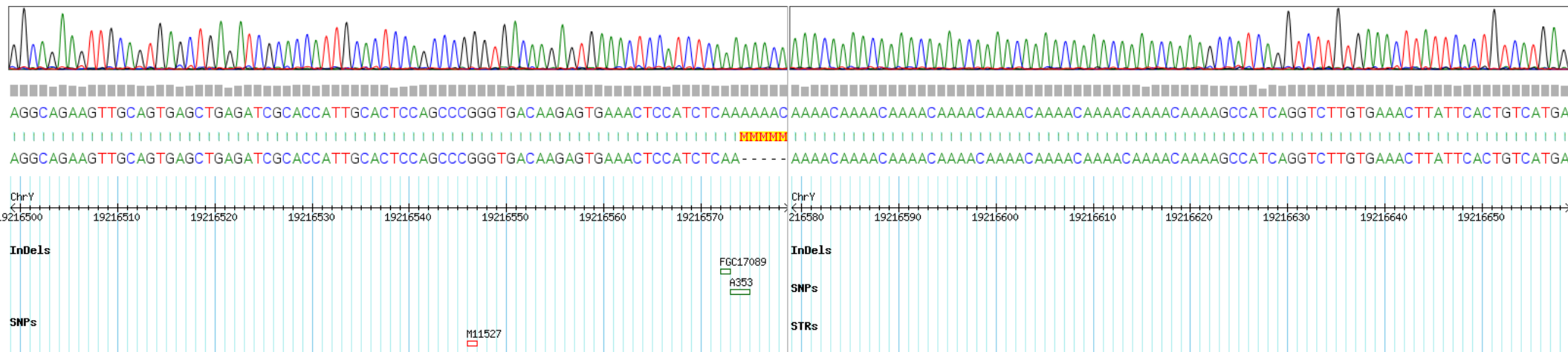
- Colorizing DNA bases helps to recognize repeat patterns
- Dot symbol for identities
- Built-in tools for alignment (Clustal W)

In the case of A353 the identities with multiple matches are not problematic
There are enough differences to exclude false mapping

Confirmation With an Independent Analysis Method

- A marker is worthless if it's only testable with one technology platform
- Ideally you want to confirm a SNP candidate with a closely related (but not identical) person in a different laboratory with a different sequencing technology
- If the SNP is available on a chip, this is an easy way to verify it.
- FTDNA and FGC both use Illumina sequencing technology, so this doesn't solve the problem. Sequencing with a different NGS method (e.g. IonTorrent or PacBio) would be acceptable
- In most cases Sanger sequencing is the most cost efficient way to crosscheck SNP candidates.

Sanger Sequencing Confirmation of A353



- Turns out that this is just an STR mutation
- A353 is likely the same as FGC17089
- Most aligners have problems with STRs
- Only a few NGS reads cover the complete STR
- Most Sequencing technologies break up the read when the STR is too long
- Brings up the question: Can STRs be tested with NGS?

STRs and NGS

YFull beta
Y-Chr Sequence Interpretation Service

Welcome, Guest! [Settings](#) [Logout](#)

Hg and SNPs
STR results
Private SNPs
Search in BAM
Check SNPs
Full Mt

Select a sample
(E1b1a1a1a) [All](#) [Submit](#)

STRs (473) [Statistic](#) [Download .CSV](#)

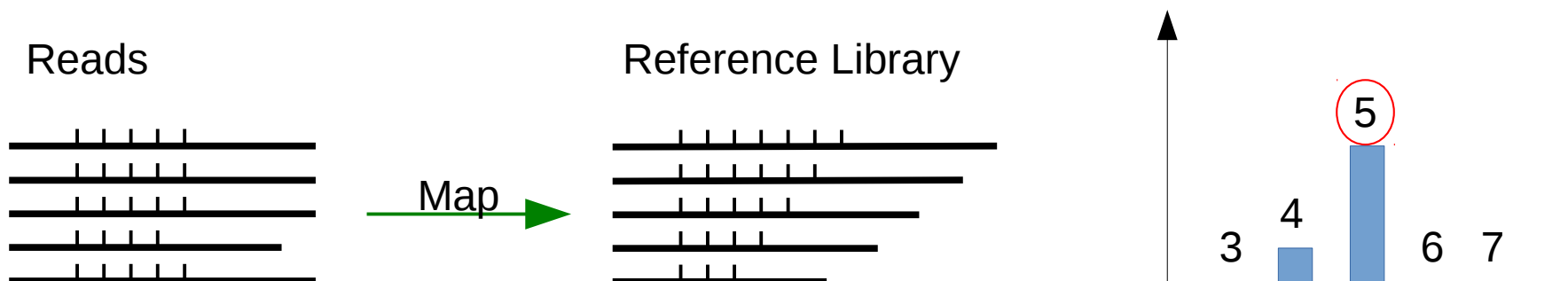
ATA71D03	CDY.1	CDY.2	DXYS156	DYF371.1	DYF371.2	DYF371.3	DYF371.4	DYF380.1	DYF380.2
11	n/a	n/a	11	10	12	13	13	10	10
DYF381.1	DYF381.2	DYF382	DYF383.1	DYF383.2	DYF384.1	DYF384.2	DYF385.1	DYF385.2	DYF386.1
8	8	12	10	11	7	8	10	10	13
DYF386.2	DYF386.3	DYF386.4	DYF387.1	DYF387.2	DYF389	DYF390	DYF391.1	DYF391.2	DYF392
13	14	14	28	32	11	10	n/a	10	8
DYF393	DYF394	DYF395.1	DYF395.2	DYF396.1	DYF396.2	DYF396.3	DYF398.2	DYF399.1	DYF399.2
n/a	8	16	16	8	8	8	8	n/a	n/a
DYF399.3	DYF400.1	DYF400.2	DYF401.1	DYF401.2	DYF403.1	DYF403.2	DYF404.1	DYF404.2	DYF405.1
n/a	20	n/a	11	15	28	29	14	14	6
DYF405.2	DYF406	DYF407.1	DYF407.2	DYF408.1	DYF408.2	DYF409.1	DYF409.2	DYF410.1	DYF410.2
12	10	11	11	13	13	12	12	10	10
DYF411.1	DYF411.2	DYF412.1	DYF412.2	DYR1	DYR10	DYR100	DYR101	DYR102	DYR103
12	12	13	13	16	10	12	12	7	10
DYR104	DYR105	DYR106	DYR107	DYR108	DYR109	DYR110	DYR111	DYR112	DYR113
11	6	8	8	9	5	5	11	12	8
DYR114	DYR115	DYR116	DYR117	DYR118	DYR119	DYR12	DYR120	DYR121.1	DYR121.2
8	30	6	10	11	n/a	10	7	10	13
DYR122.1	DYR122.2	DYR123	DYR124.1	DYR124.2	DYR124.3	DYR125.1	DYR125.2	DYR126	DYR127
7	7	14	10	12	12	25	26	13	9
DYR128.1	DYR128.2	DYR129	DYR13	DYR130	DYR131	DYR14	DYR15	DYR17.1	DYR17.2
9	12	13	10	13	7	11	11	11	11
DYR17.3	DYR18.1	DYR18.2	DYR19	DYR2	DYR20	DYR23	DYR26	DYR27	DYR28
12	12	12	0	14	11	12	12	10	12

<http://www.yfull.com>

Apparently it kind of works... However especially long STRs have problems.

An Alternative Approach for Scoring STRs

- Yfull likely uses Tandem Repeats Finder (TRF) on the final consensus sequence.
- However many mappers only consider the reference sequence as the only truth. => Repeat counts that differ significantly from the repeat count of the reference will be entirely rejected. They will not contribute to the consensus sequence.
- A better approach is to map the raw reads to a library of short reference sequences (for each possible repeat count a separate one). Wherever you get the highest number of mapped reads will indicate the correct repeat count

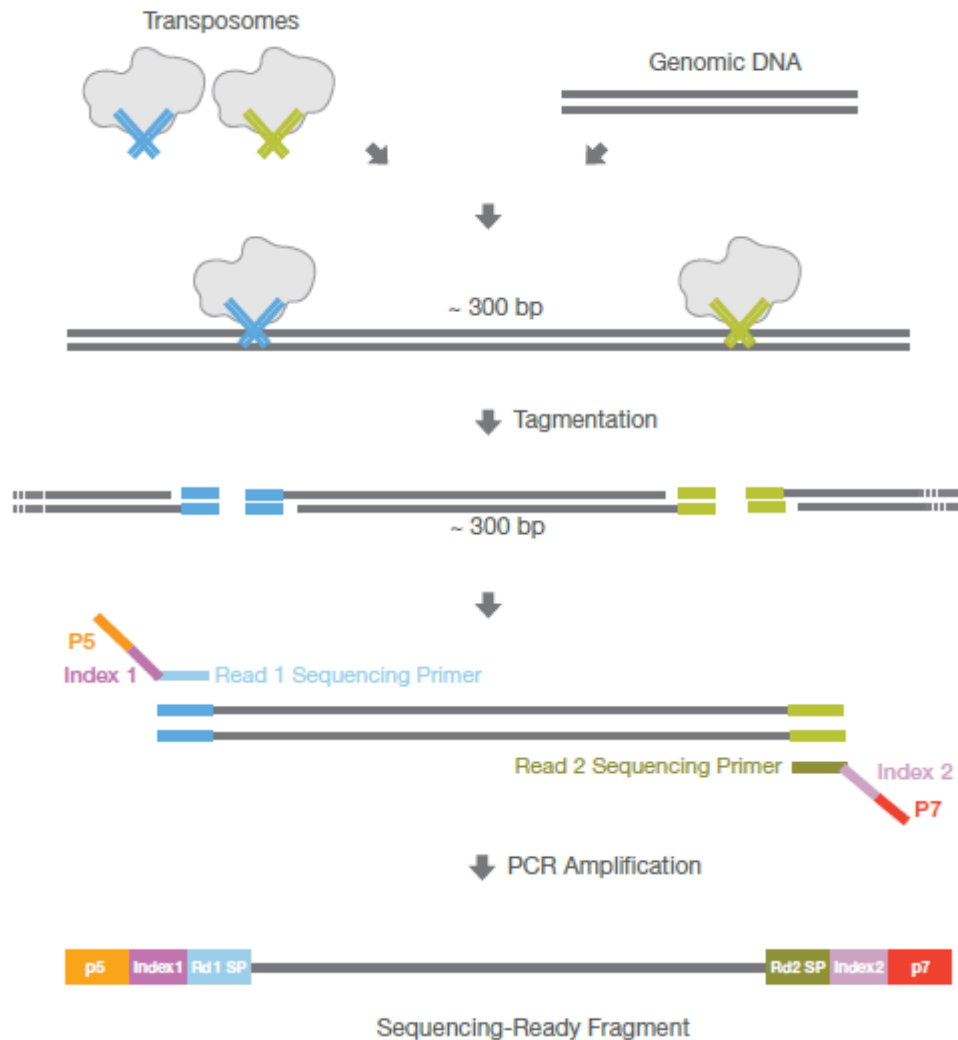


How Was the BigY Enrichment Assay Designed?

- Illumina has a “proprietary” process to generate 80 base long enrichment probes for the regions that their customers provide (in a BED file)
- They really just “tile” the probes in equal distance. All they care about is to get the G+C ratio of the probes in the ballpark of 48%
- I wanted high Y chromosome specificity, therefore I asked them to space the probes closer and increase the probe frequency
- Then I applied filters by BLASTing each probe automatically against the full genome reference sequence *EXCEPT* the Y. Probes that were matches elsewhere were removed from the pool.
- Probes for known markers that were on the Y tree at that time were forced to be added to the pool to ensure coverage and backward compatibility

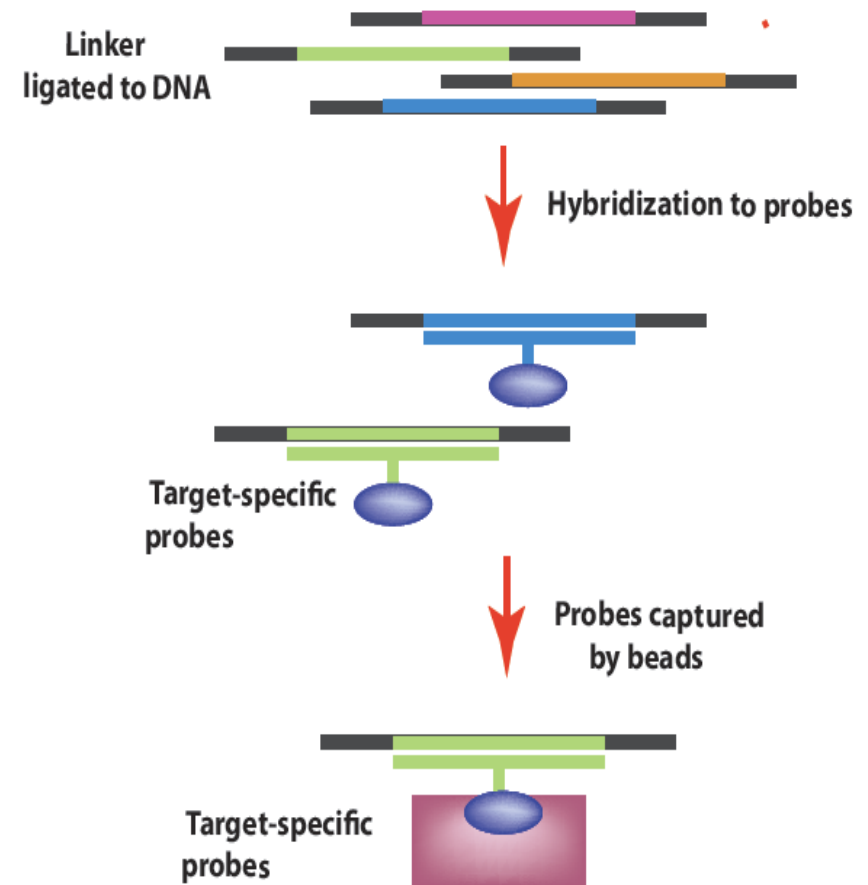
Illumina Enriched Library Sequencing

Nextera Sample Preparation Biochemistry



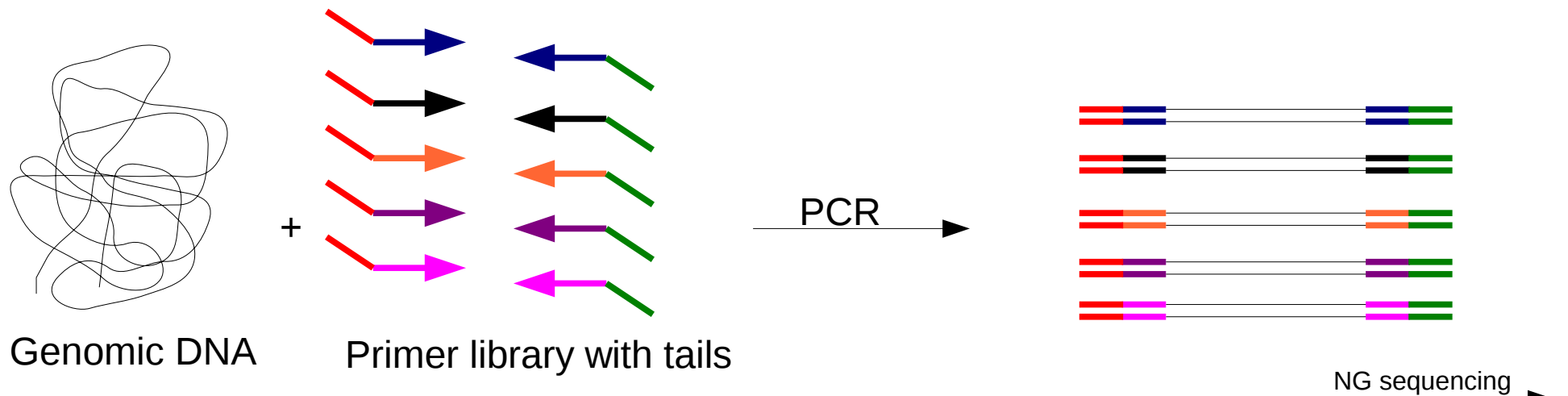
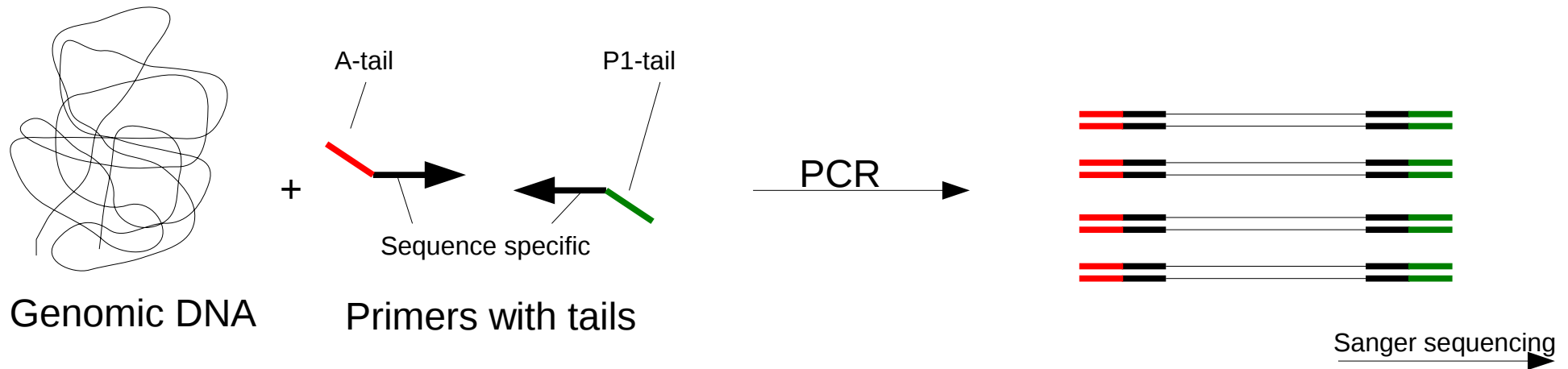
Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indices to each fragment.

In-Solution Capture



Source: Illumina / Wikipedia

Alternative: Direct PCR Library



Use the same primers for Sanger & NGS!

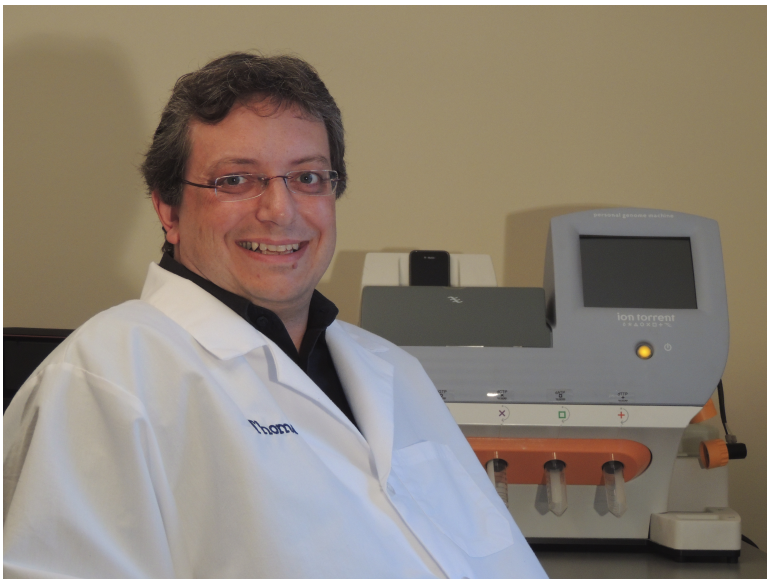
The YSEQ Team



Astrid



Heather



Thomas

<http://YSEQ.net>