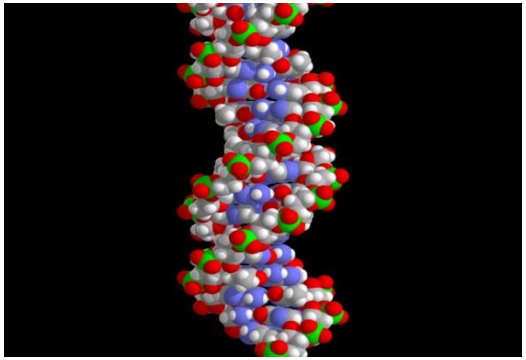
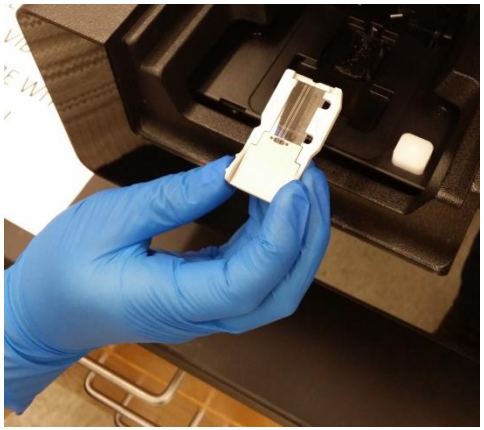
UNT | HEALTH SCIENCE CENTER™

Overview of Moving Implementation Mountains: Forensics, NGS, Bioinformatics

ISHI Workshop 2016





Bruce Budowle

Center for Human Identification

University of North Texas Health Science Center

Fort Worth, Texas USA

Issues/Goals

- What is MPS?
- What is the technology?
- What is the workflow?
- Markers
- Applications
- Bioinformatics
  - Challenge of data analysis
  - Interpretation

The Human Genome Project

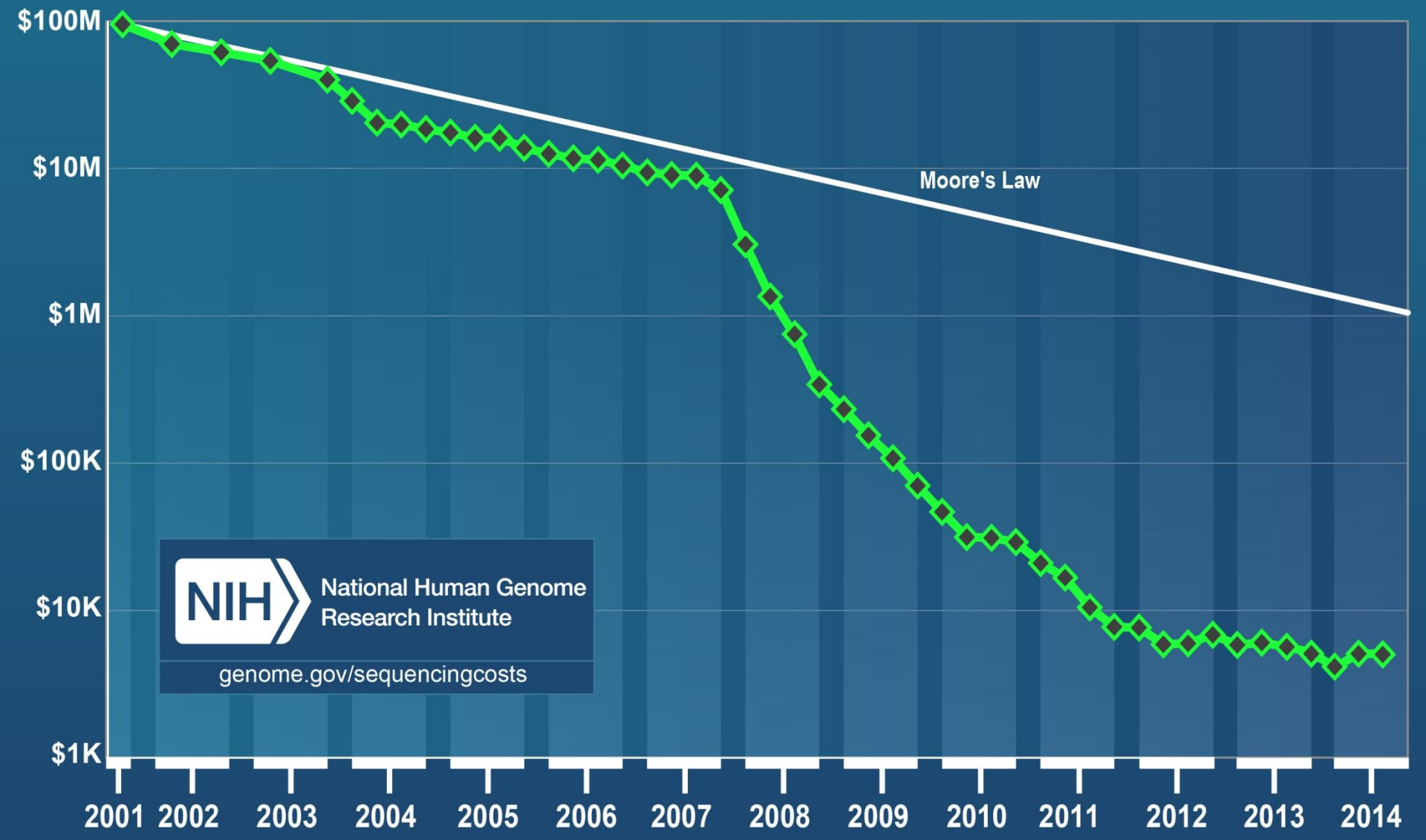
Scale

1990-2003

- 13 years
- ~40 Institutions
- 8-9X Coverage
- $3.8 Billion

Human Genome Project - 10 years and cost $2.7 billion

Cost per Genome



Personalized Machine Sequencing

Genome Center Capabilities in the Forensic Laboratory

http://www.biomatrica.com/strboost.php

illumina

Roche

life technologies

PACIFIC BIOSCIENCES

Helicos BioSciences Corporation

ion torrent

Oxford NANOPORE Technologies

HiSeq2000 / 2500

Ion torrent

Ion Proton

GS-FLX

MiSeq

MPS In Forensic DNA Typing

- Areas of Interest
  - STRs
  - Mitochondria
  - HID SNPs
  - Ancestry SNPs
  - Phenotype SNPs
  - Pharmacogenetics
  - Microbial Forensics
  - Animal and Plant Forensics

MPS for Forensic DNA Typing

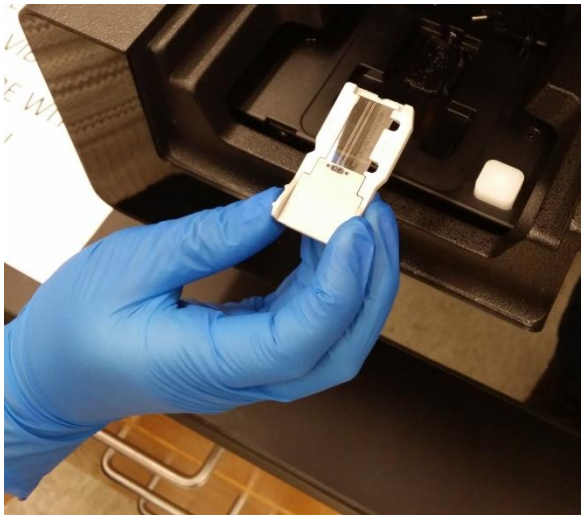
- Addition to current capabilities!
  - More markers
  - Different markers
  - Throughput
  - Identity testing
  - Kinship analysis
  - Mixture interpretation
  - Reduction of artifacts

Not a Paradigm Shift!

- Same principles
- Similar molecular biology
- Same and similar markers
- Similar applications
- Adjunct technology
- Consider verification/validation
- Front end through PCR is the same
- Current experience will help with back end – interpretation
- Just typing on steroids!

MPS

- NGS is no longer next generation, consider:
  - Massively Parallel Sequencing (MPS)

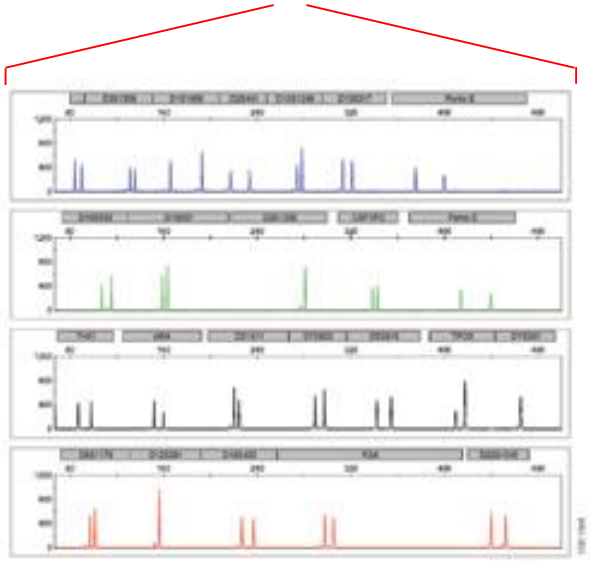


MPS and Forensic STR"Gold Standard"

- Capillary electrophoresis fragments (size differences)

Targeted MPS

~13-24 markers



Increasing size (bp)

100s of Markers

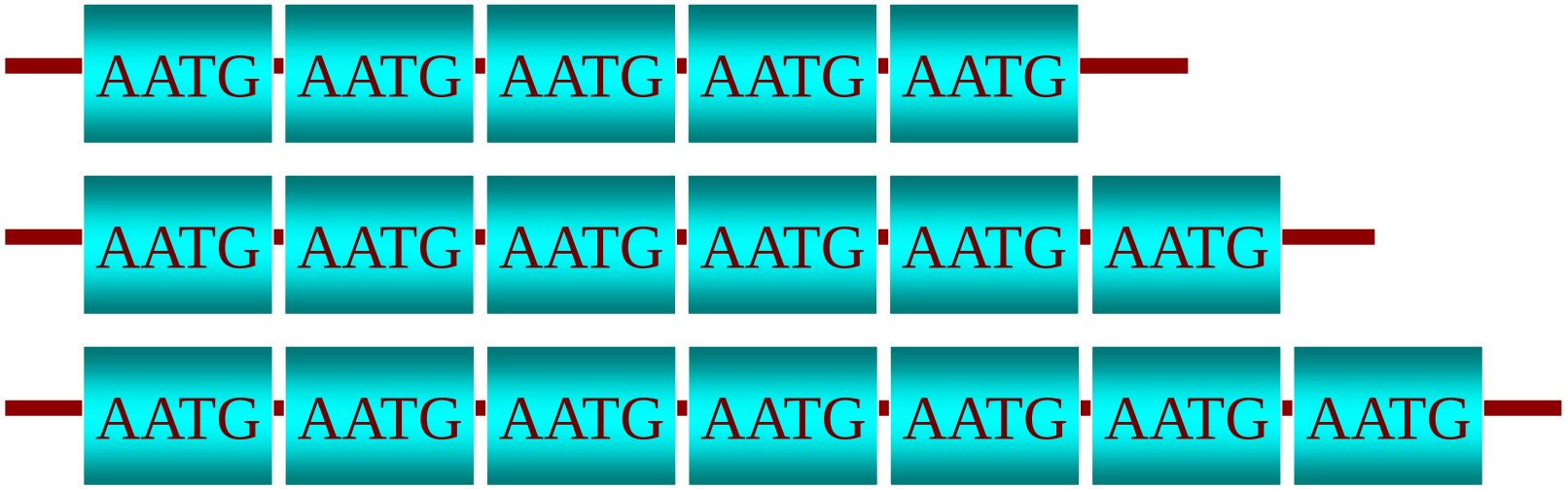
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG
GTGTGATGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTG

Massively parallel sequencing
Fragments can overlap in size
Read counts / noise

Length & Sequence Variations

STRs

- Current mainstay for identity testing
- High discrimination power



Types of SNPs

- Individual Identification SNPs:
  - SNPs that collectively give very low probabilities of two individuals having the same multisite genotype; individualization, High heterozygosity, low Fst
- Ancestry Informative SNPs:
  - SNPs that collectively give a high probability of an individual's ancestry being from one part of the world or being derived from two or more areas of the world
- Lineage Informative SNPs:
  - Sets of tightly linked SNPs that function as multiallelic markers that can serve to identify relatives with higher probabilities than simple di-allelic SNPs
- Phenotype Informative SNPs:
  - SNPs that provide high probability that the individual has particular phenotypes, such as a particular skin color, hair color, eye color, etc.
- Pharmacogenetic SNPs – molecular autopsy

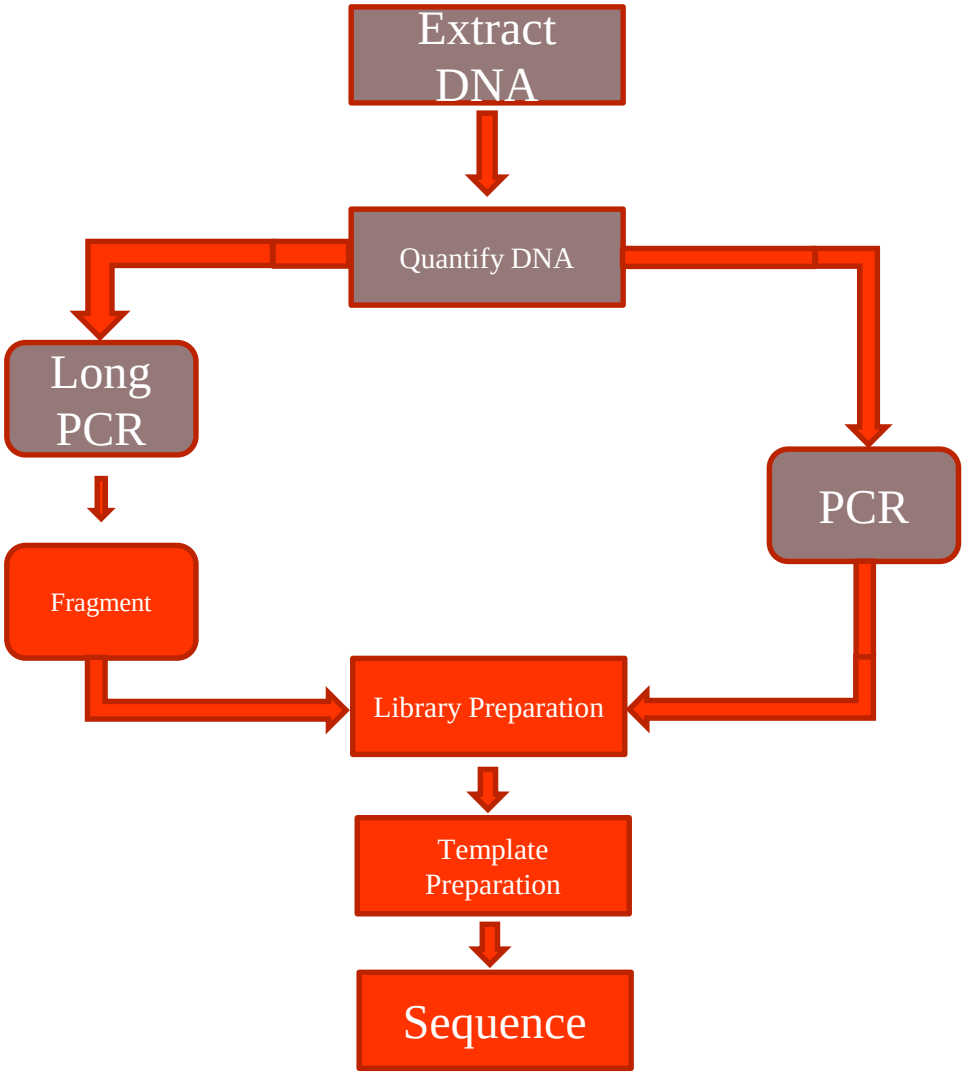
Forensically relevant SNP classes

Bruce Budowle[1] and Angela van Daal[2]

BioTechniques 44:603-610 (25th Anniversary Issue, April 2008)
doi 10.2144/000112806

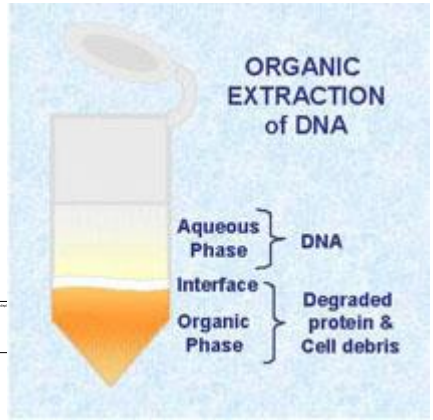
Forensic samples that contain too little template DNA or are too degraded require alternate genetic marker analyses or approaches to what is currently used for routine casework. Single nucleotide polymorphisms (SNPs) offer promise to support forensic DNA

Generic MPS Workflow



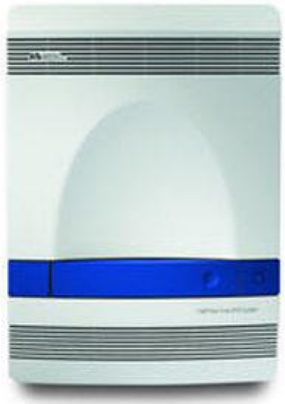
Same DNA extraction methods currently used!

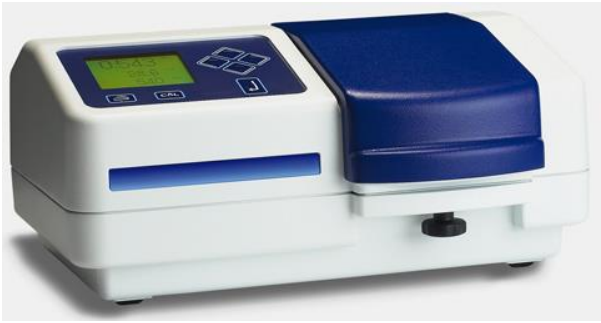
- Nuclear DNA
- Mitochondrial DNA



Same DNA quantitation methods currently used!

Qubit may be new
but just a fluorometer





Library Preparation

- Getting the DNA ready for sequencing

1. Genomic DNA (Nuclear/Mitochondrial)
2. Get the DNA to the right size
3. Add adapters, priming sites, and barcodes
4. Normalize and pool libraries
5. Cloning

Library Preparation

Adapters:

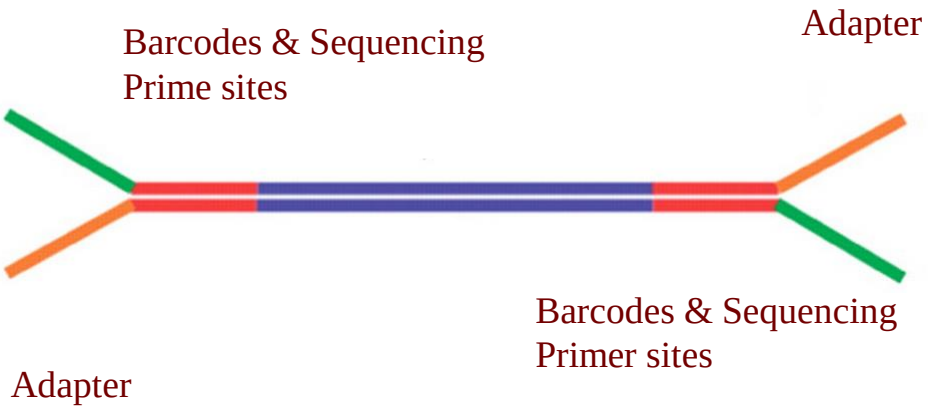
short pieces of DNA for DNA fragment to attach to a solid support for clonal amplification

Barcodes:

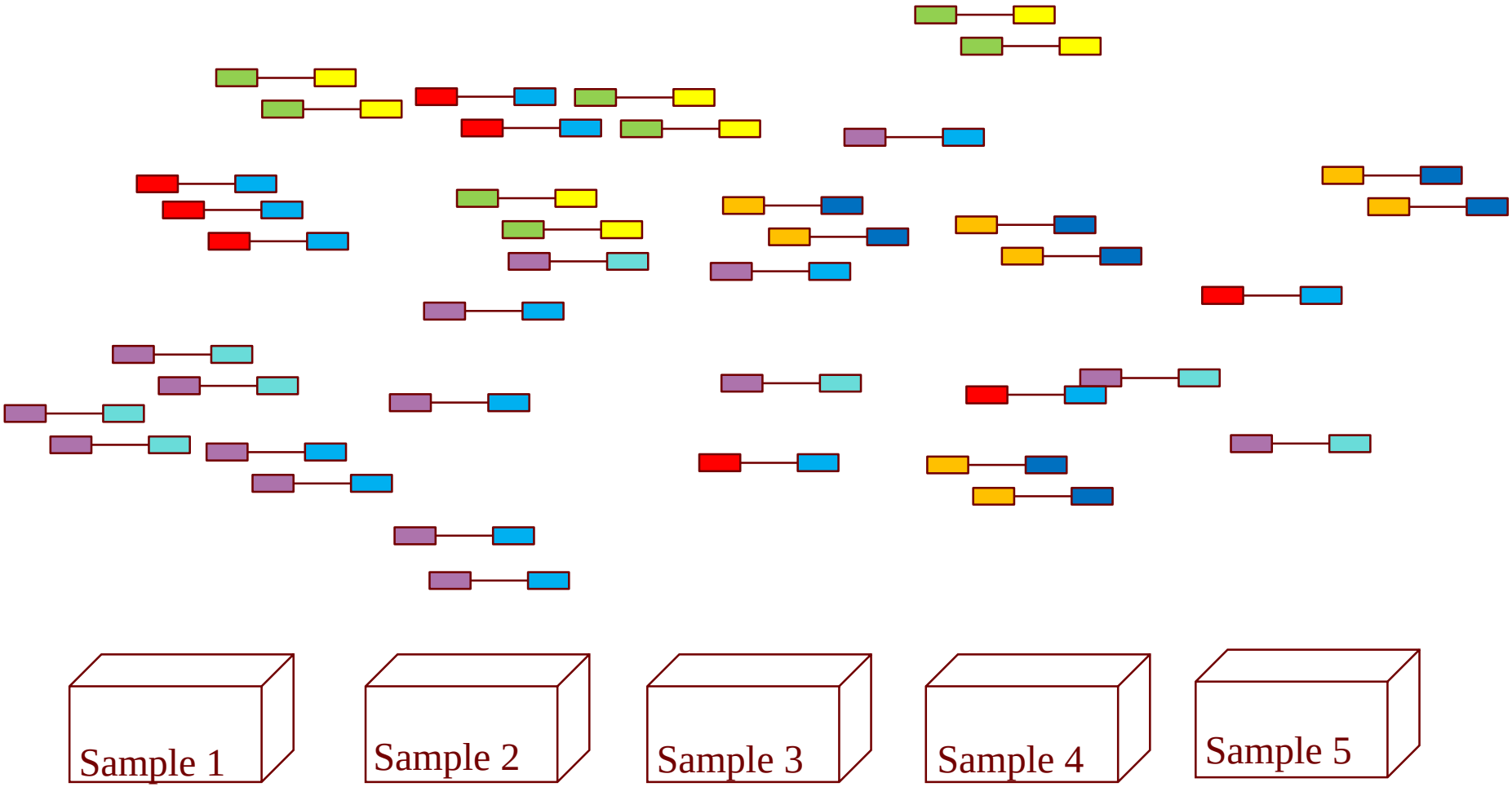
short pieces of DNA necessary for identifying multiplexed samples bioinformatically

Priming sites:

Short sequences for primers to bind to initiate sequencing by synthesis

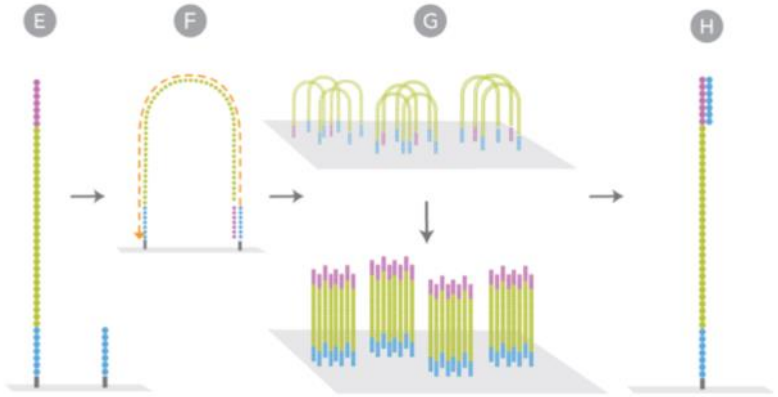


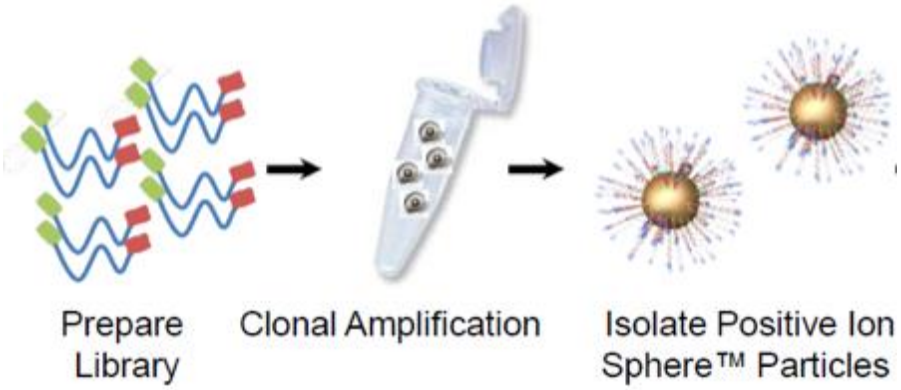
Multiplexing/Demultiplexing Barcoding/Index Sequences



Amplification of Library Fragments

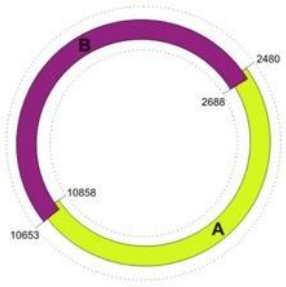
- Bridge Amplification
- Emulsion PCR

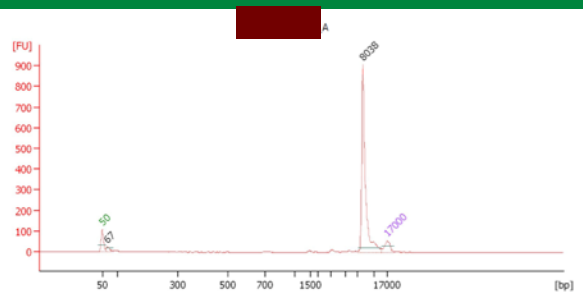




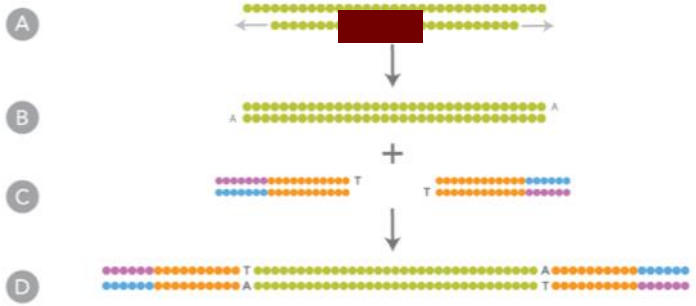
Instruments

Illumina MiSeq Workflow



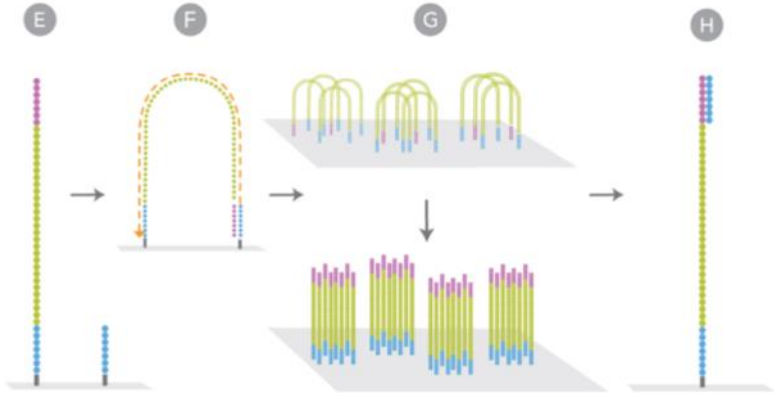


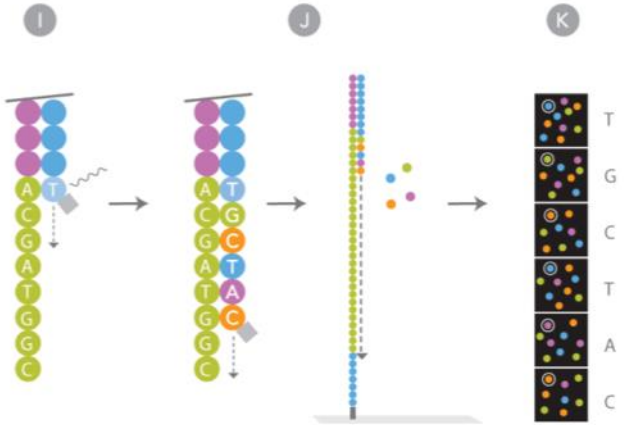
PCR



Nextera XT

Flowcell



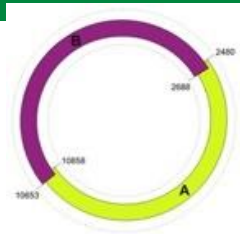


Cluster Generation & Sequencing by Synthesis

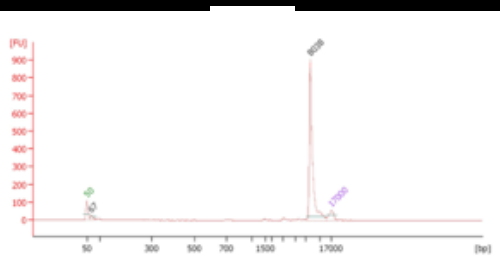
Illumina MiSeq

Modified and by courtesy of Illumina and W Parson

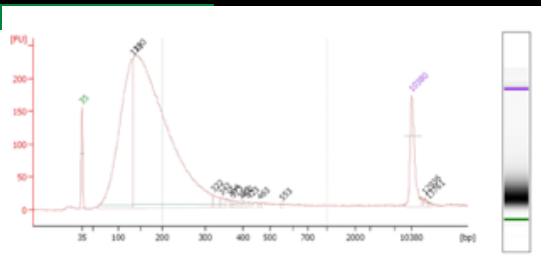
PGM/Ion Torrent Workflow



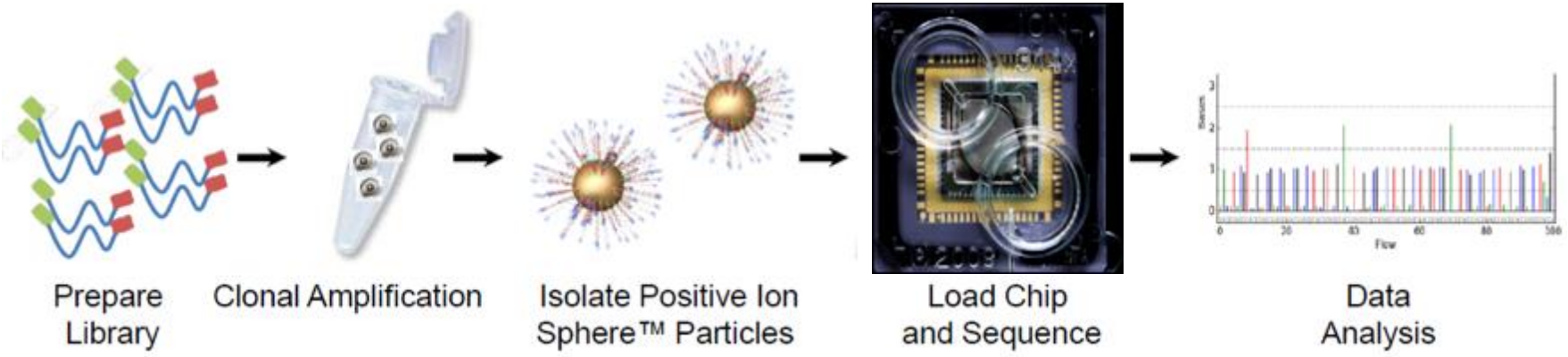
Parson et al 2013



PCR



e-shearing



Ion Chef

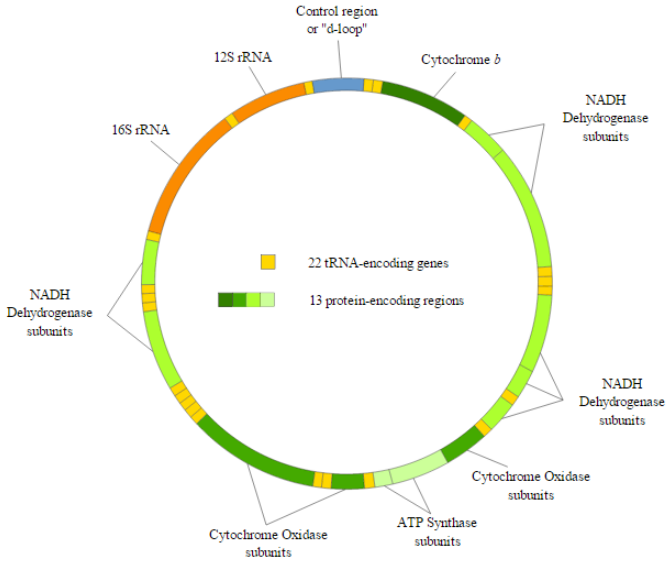
PGM™

Torrent Server and Torrent Browser

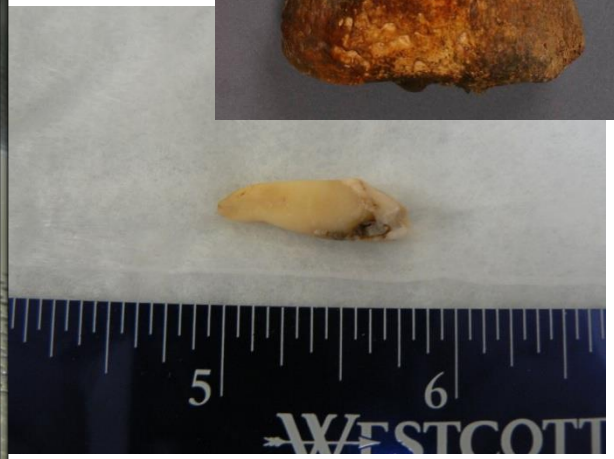
Modified and by courtesy of Thermo Fisher Scientific and W Parson

Practical Applications

- Mitochondrial DNA Sequencing
  - mtGenome (~16,569 bp)
    - Long PCR/Short PCR



Analysis of Difficult Samples



mtDNA is the most successful marker

Advantages of mtDNA Analysis

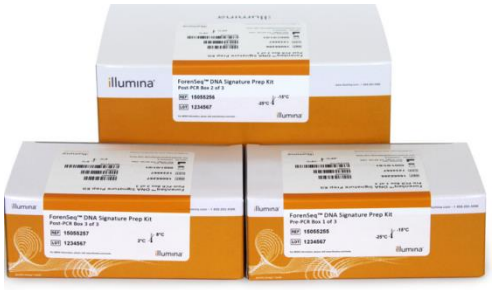
- High copy number
  - limited sample
  - hair, teeth, bones
- Less prone to degradation
  - structure, location
- Maternal inheritance
  - maternal relatives source of known sample in missing persons cases
- Highly variable among individuals

Limitations of Sanger Sequencing

1) Typing procedure currently used for mtDNA is labor intensive
2) Focus usually on HV1 and HV2
3) Analysis of results is time consuming
4) Costly (prices range from $1000 to $3000 per sample)
5) Variation in intensity of peaks
6) Not quantitative– impacts mixture interpretation
7) Heteroplasmy difficulties

Practical Applications

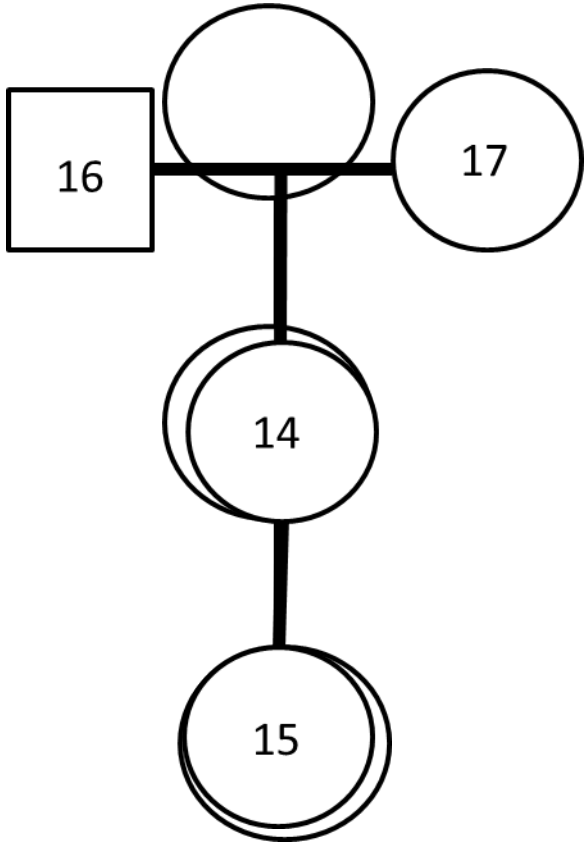
- SNPs/INDELs and STRs
  - Casework/Databasing
    - PCR
      - ForenSeq (Illumina)
      - AmpliSeq (Life Tech)
      - PowerSeq (Promega)



Promega

Identifying Relationships

Genotypes from STRs and Identity SNPs allow for expansion and refinement of the partial pedigree identified with the mitochondrial haplotypes



Identifying Relationships

STRs:

Likelihood Ratio Results:

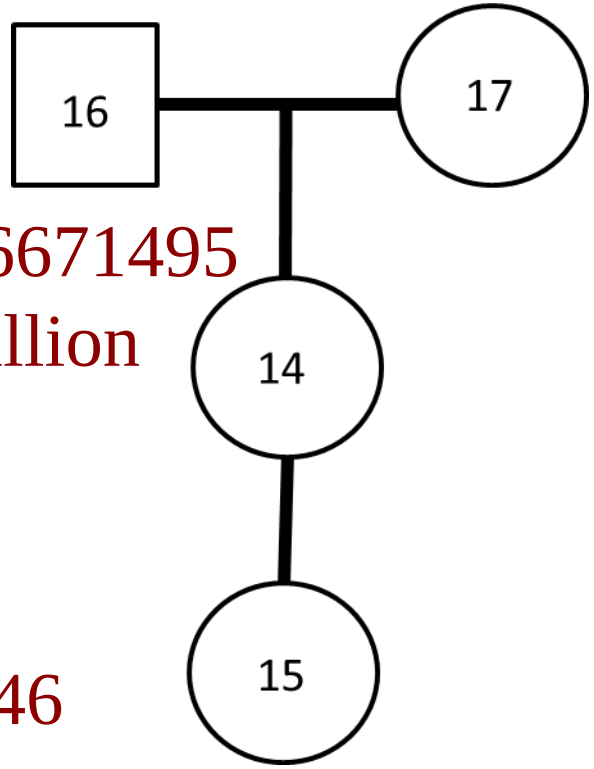
Posterior probability = 0.999999996671495

Combined likelihood ratio = 300 million

SNPs:

Likelihood Ratio Results:

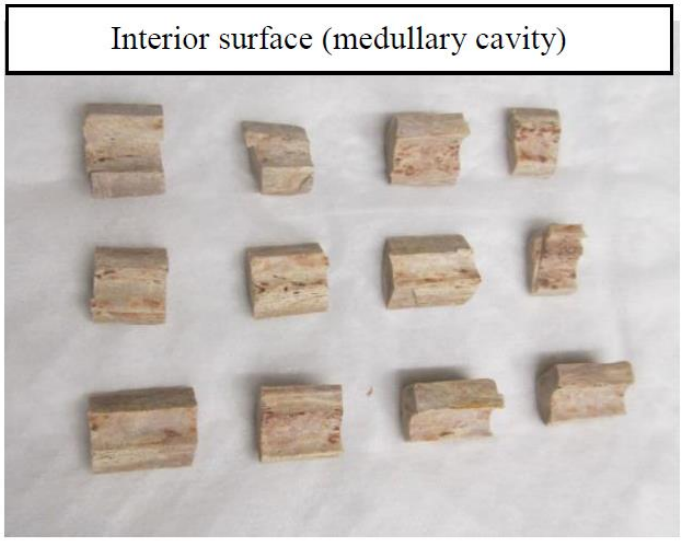
Combined likelihood ratio = 3.34 E46

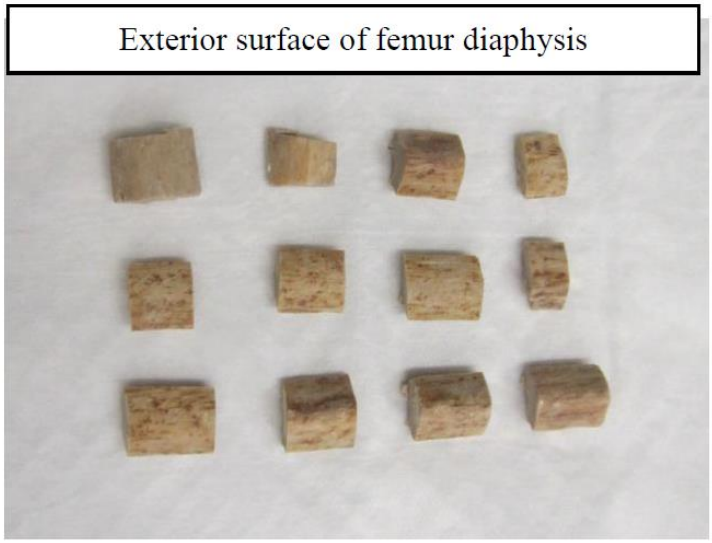


Internal ThermoFisher kinship algorithm used for calculations

Bone Sections

- Right femur
- Diaphysis surface-sanded w/Dremel® 4000 Rotary Tool and sterile grinding stone
- Sectioned with Stryker® autopsy saw





Consensus Profile

	Loci Typed						
	CE	ForenSeq					
	Y-STRs	Y-STRs	Autosomal STRs	X-STRs	HID SNPs	AIMs	Phenotypic
Deadwood FEMUR 008.001 E1	16	10	22	2	84	50	24
Deadwood FEMUR 008.002 E1	15	12	23	3	87	50	24
Deadwood FEMUR 007.001 E1	6	2	13	0	40	23	12
Deadwood FEMUR 008.002 E2	13	12	26	4	71	40	17

Only alleles over 30x coverage are listed

Haplogroup Predictor

Yfiler
Results Table

Haplo-group	Fitness score	Proba-bility (%)
E1a	2	0.0
E1b1a	4	0.0
E1b1b	5	0.0
G1	3	0.0
G2a	3	0.0
G2b	1	0.0
I1	2	0.0
I2a	10	0.0
I2b	2	0.0
J1	3	0.0
J2a	5	0.0
J2b	1	0.0
L	7	0.0
N	7	0.0
O	10	0.0
Q	10	0.0
R1a	11	0.0
R1b	29	100.0
R2	6	0.0
T	5	0.0

ForenSeq
Results Table

Haplo-group	Fitness score	Proba-bility (%)
E1a	2	0.0
E1b1a	29	0.0
E1b1b	18	0.0
G1	8	0.0
G2a	11	0.0
G2b	0	0.0
I1	6	0.0
I2a	35	0.0
I2b	10	0.0
J1	11	0.0
J2a	27	0.0
J2b	9	0.0
L	21	0.0
N	13	0.0
O	26	0.0
Q	26	0.0
R1a	31	0.0
R1b	79	100.0
R2	9	0.0
T	17	0.0

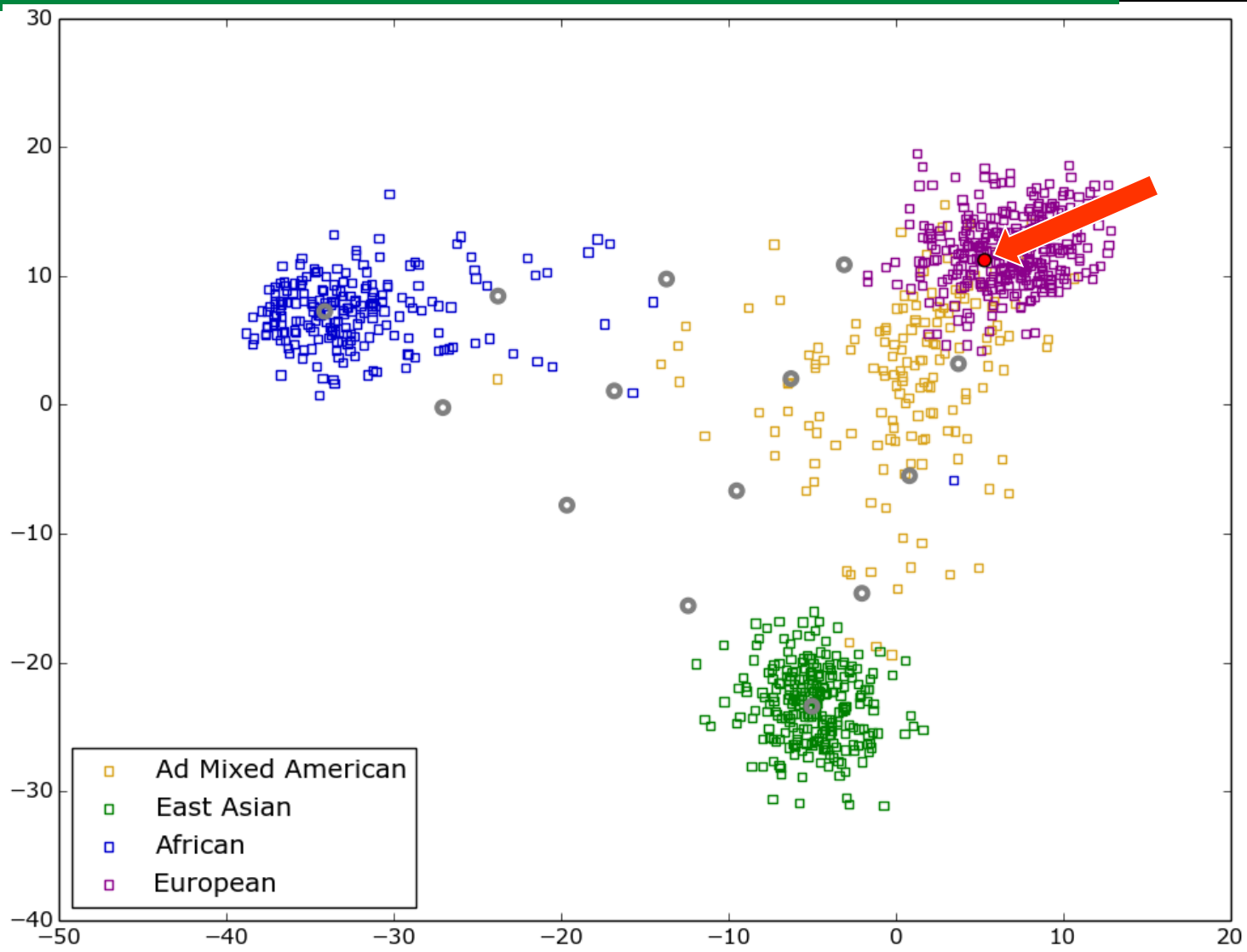
Combined
Results Table

Haplo-group	Fitness score	Proba-bility (%)
E1a	1	0.0
E1b1a	8	0.0
E1b1b	9	0.0
G1	4	0.0
G2a	6	0.0
G2b	0	0.0
I1	3	0.0
I2a	18	0.0
I2b	4	0.0
J1	5	0.0
J2a	10	0.0
J2b	3	0.0
L	9	0.0
N	7	0.0
O	14	0.0
Q	14	0.0
R1a	17	0.0
R1b	40	100.0
R2	5	0.0
T	10	0.0

http://www.hprg.com/hapest5/

European Ancestry

ForenSeq Aims



Phenotype Probabilities

HIrisPlex hair & eye colour prediction tool

	SNP ID		Minor allele	No.
1	N29insA	1	A	0
2	rs11547464	2	A	0
3	rs885479	3	T	0
4	rs1805008	4	T	2
5	rs1805005	5	T	0
6	rs1805006	6	A	0
7	rs1805007	7	T	0
8	rs1805009	8	C	0
9	Y152OCH	9	A	0
10	rs2228479	10	A	0
11	rs1110400	11	C	0
12	rs28777	12	C	0
13	rs16891982	13	C	0
14	rs12821256	14	G	0
15	rs4959270	15	A	1
16	rs12203592	16	T	1
17	rs1042602	17	T	0
18	rs1800407	18	A	0
19	rs2402130	19	G	0
20	rs12913832	20	T	1
21	rs2378249	21	C	1
22	rs12896399	22	G	2
23	rs1393350	23	T	0
24	rs683	24	G	1

HIrisPlex: Walsh et al. (2013)

Hair color	
Brown	0.19
Red	0.69
Black	0.04
Blond	0.09

Hair Color Shade	
Light	0.71
Dark	0.29

Eye Color	
Brown	0.51
Intermediate	0.22
Blue	0.27

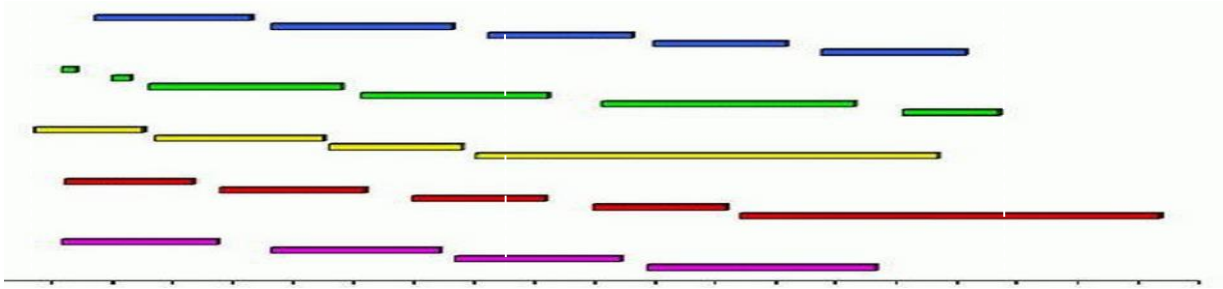
http://hirisplex.erasmusmc.nl
Accessed on 10-07-2014

Back to STRs

- Options
  - Ion Torrent Panels
  - ForenSeq Panel
  - PowerSeq™ Kit
- All Point to added value with MPS
  - Diversity
  - Mixture
  - Stutter
  - Dynamic range

MPS and miniSTR Primer Design

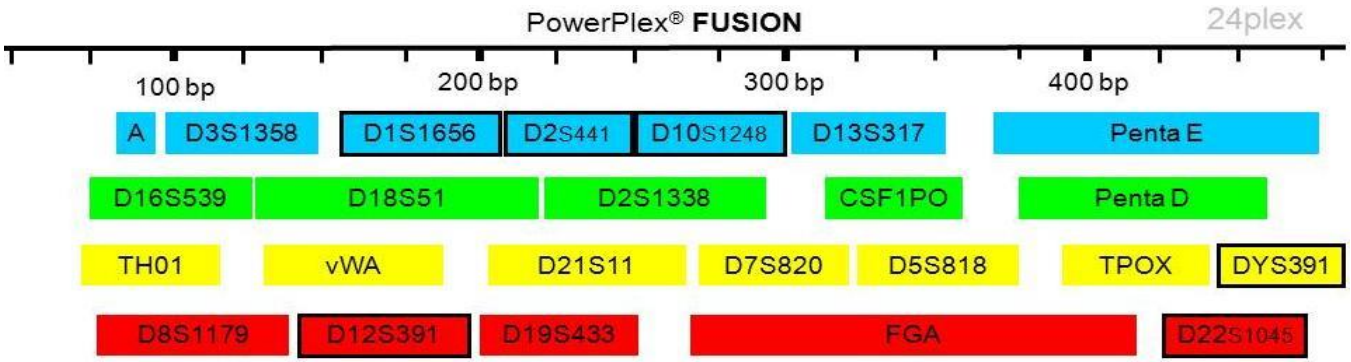
CE



NGS

Evaluation of the Prototype PowerSeq™ Auto System

Prototype PowerSeq™ Auto system is a subset of the MPS version of PowerPlex™ Fusion kit



Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing

Xiangpei Zeng[a], Jonathan L. King[a], Monika Stoljarova[a], David H. Warshauer[a], Bobby L. LaRue[a], Antti Sajantila[a,b], Jaynish Patel[c], Douglas R. Storts[c], Bruce Budowle[a,d,*]

Early Access Ion AmpliSeq GlobalFiler Mixture ID Panel

- Amplifies 113 markers
- Multiplex of four different marker types
  - STRs
  - SNPs
  - Indels
  - Microhaplotypes

Table 3. Precision ID GlobalFiler NGS STR kit markers.

Locus	Repeat Structure	Source	Chr
TPOX	AATG	CODIS	2
D3S1358	TCTA/TCTG	CODIS	3
FGA	CTTT/TTCC	CODIS	4
CSF1PO	AGAT	CODIS	5
D5S818	AGAT	CODIS	5
D7S820	GATA	CODIS	7
D8S1179	TCTA/TCTG	CODIS	8
TH01	TCAT	CODIS	11
vWA	TCTA/TCTG	CODIS	12
D13S317	TATC	CODIS	13
D16S539	GATA	CODIS	16
D18S51	AGAA	CODIS	18
D21S11	TCTA/TCTG	CODIS	21
AMEL-X	NA	Sex determination	X
AMEL-Y	NA	Sex determination	Y
rs2032678	NA	Sex determination	Y
D1S1656	TAGA	Expanded CODIS	1
D2S441	TCTA/TCAA	Expanded CODIS	2
D2S1338	TGCC/TTCC	Expanded CODIS	2
D10S1248	GGAA	Expanded CODIS	10
D12S391	AGAT/AGAC	Expanded CODIS	12
D19S433	AAGG/TAGG	Expanded CODIS	19
D22S1045	ATT	Expanded CODIS	22
DYS391	TCTA	Expanded CODIS	Y
D1S1677	TTCC	MPS	1
D2S1776	AGAT	MPS	2
D3S4529	ATCT	MPS	3
D4S2408	ATCT	MPS	4
D5S2800	GATA/GATT	MPS	5
D6S474	GATA/GACA	MPS	6
D6S1043	AGAT/AGAC	MPS	6
D12ATA63	TAA/CAA	MPS	12
D14S1434	CTGT/CTAT	MPS	14

Precision ID GlobalFiler NGS STR Panel

Analyze DNA mixtures more efficiently

The Applied Biosystems™ Precision ID NGS System for human identification can help you solve tough cases by getting more information from your challenging samples. Adopting next-generation sequencing (NGS) in your laboratory is now simpler than ever. Applied Biosystems™ Precision ID GlobalFiler™ NGS STR

Simplicity

- As little as 1 ng of input DNA
- Interpret variants with easy-to-use HID STR Genotyper plug-in

ForenSeq DNA Signature Prep Kit

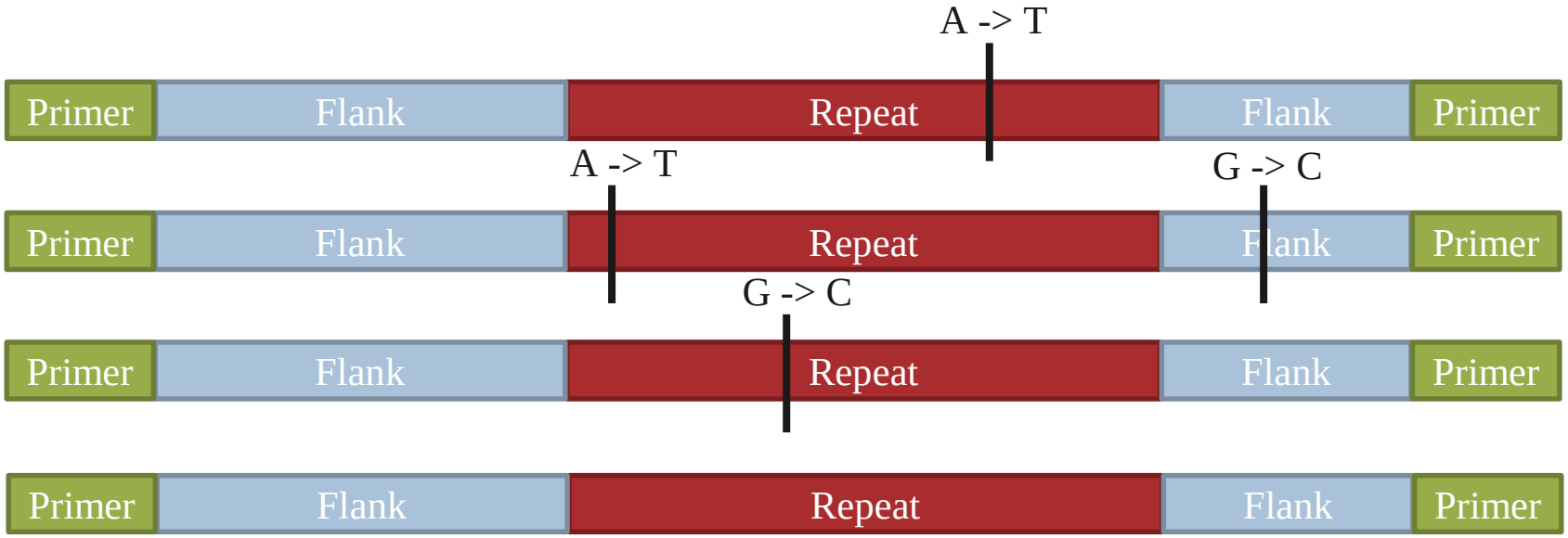
Table 1: ForenSeq DNA Signature Prep Kit—Forensic Loci

Feature	Number of Markers[a]	Amplicon Size Range (bp)	Included in DNA Primer Mix A	Included in DNA Primer Mix B[b]
Global Autosomal STRs	27	61–467	Yes	Yes
Y-STRs	24	119–390	Yes	Yes
X-STRs	7	157–462	Yes	Yes
Identity SNPs	94	63–231	Yes	Yes
Phenotypic SNPs	22	73–227	No	Yes
Biogeographical Ancestry SNPs	56	67–200	No	Yes

Figure 2: ForenSeq DNA Signature Prep Kit—The ForenSeq DNA Signature Prep Kit includes all reagents required to prepare 384 DNA libraries for sequencing, including PCR reagents, index adapters, and purification and normalization beads.

STR Sequence Variation

- MPS provides nucleotide level data for high resolution genotyping



- If ignore the deeper information, can generate size-based allele calling

Variation Example at D3S1358 Locus

Stutter sequence for 15 alleles

AGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGAT

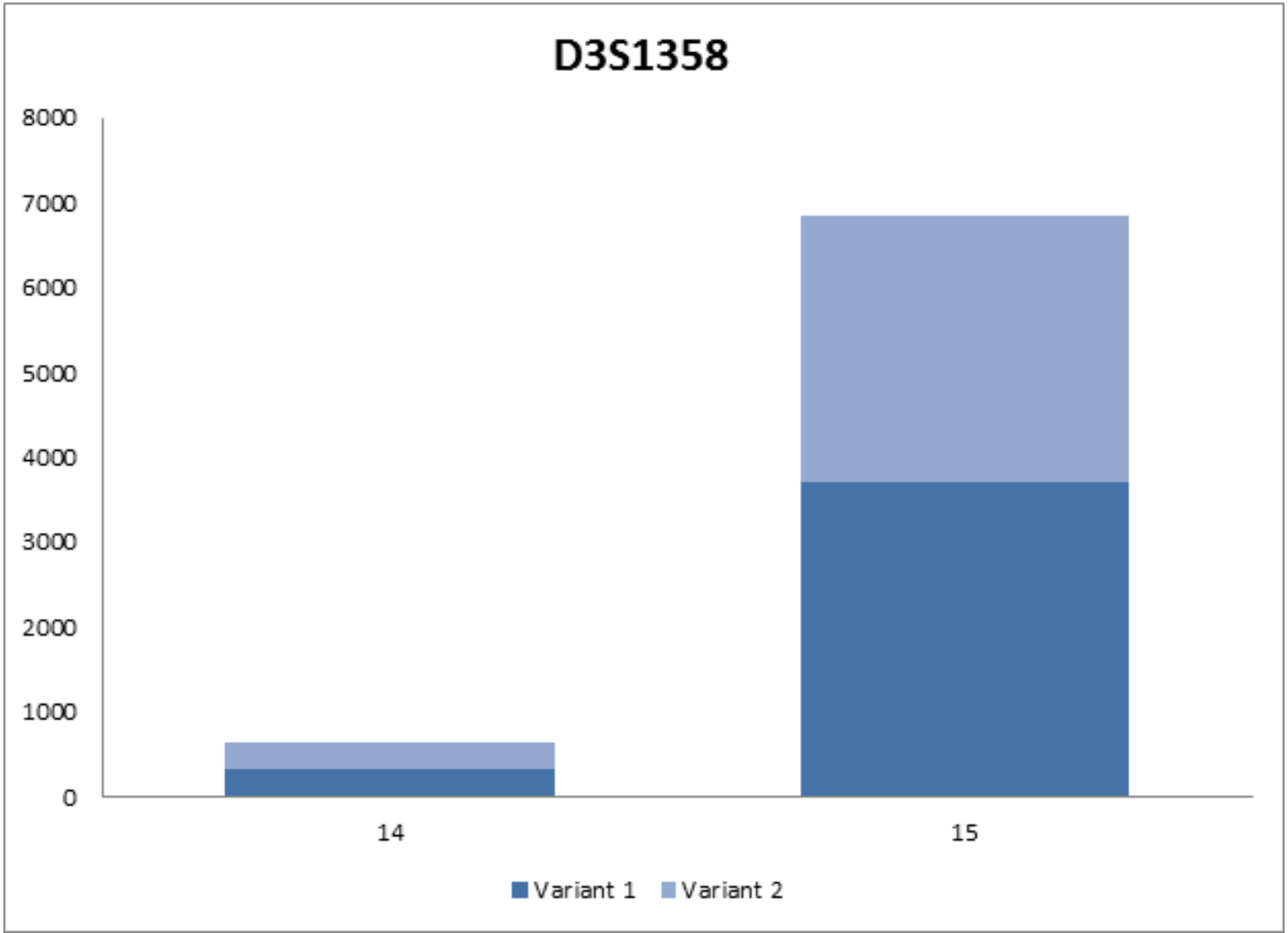
AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGATAGAT

Sequence for 15 alleles

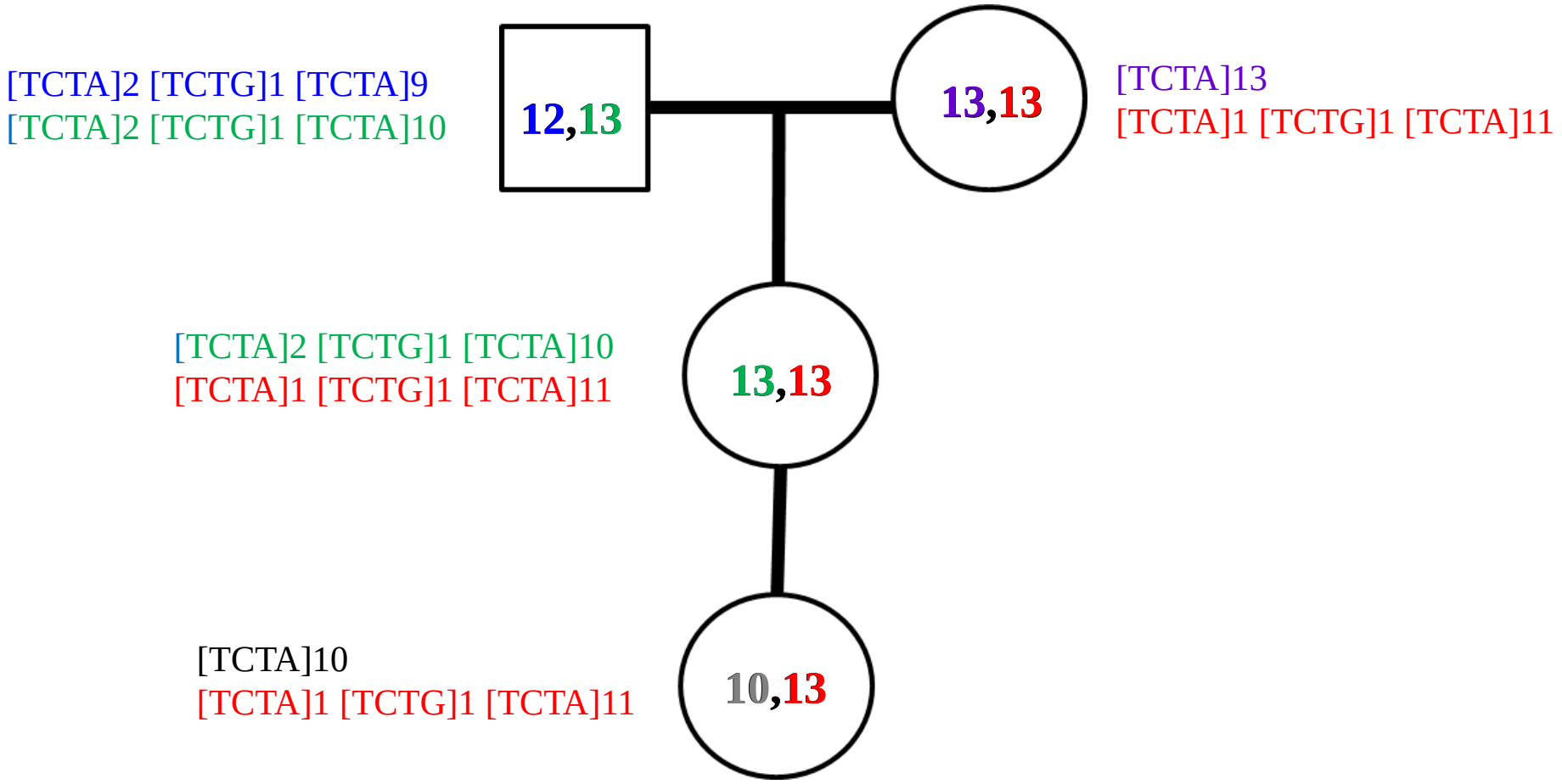
AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGAT

AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGATAGAT

Allele and Stutter Distribution for D3S1358 Homozygote

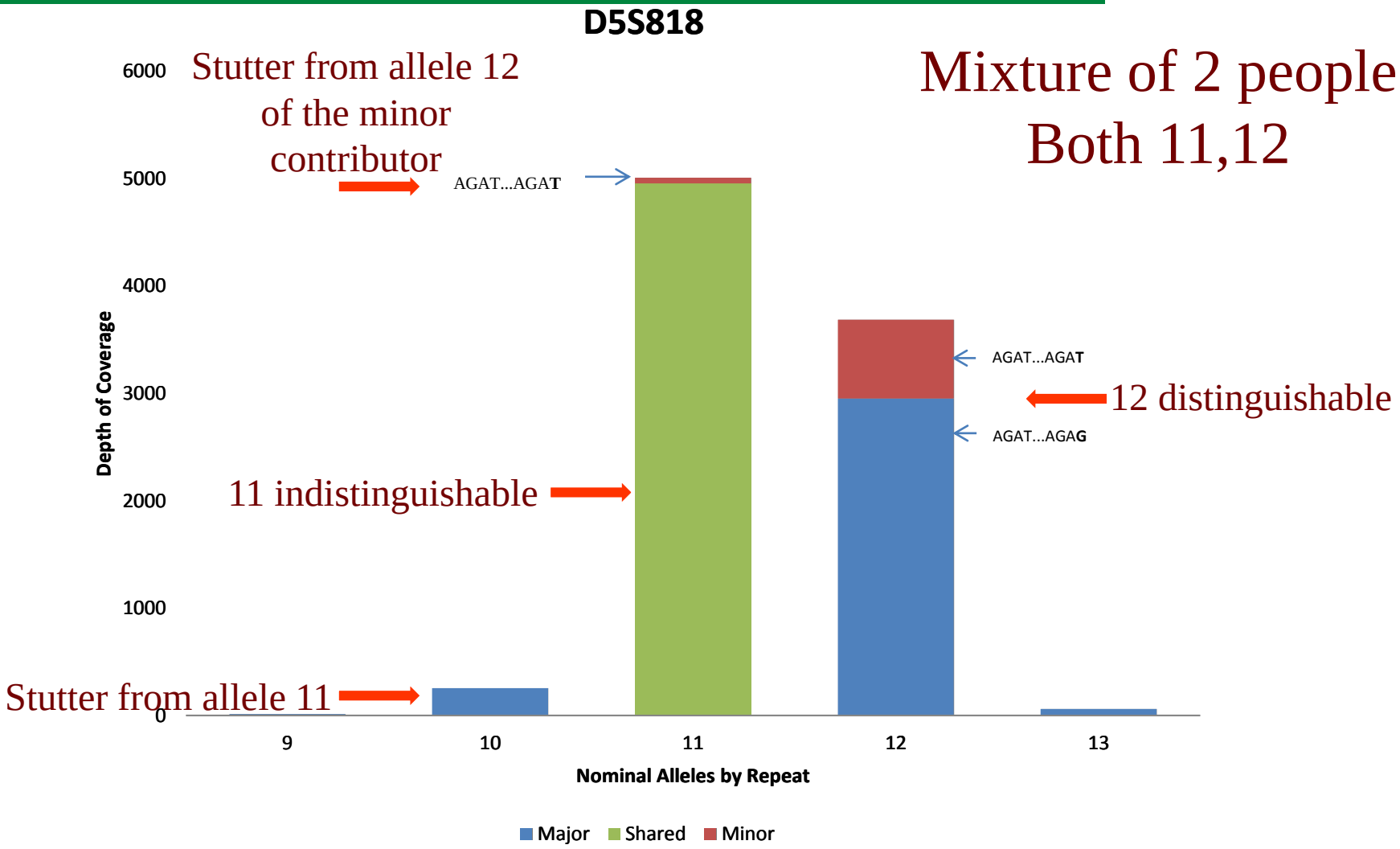


Identifying Relationships



Sequence variants present within the D8S1179 alleles

Minor and Major Contributor Alleles



Bioinformatics

- Unprecedented access to biological data
  - data acquisition
- Managing biological databanks with numerous contributors and users
  - store, organize, networks
- Extracting useful information from large and dense biological data
  - manipulate, visualize
- Assembling molecular pieces into predictive models of biological systems for *in silico* experiments
  - modeling, inference
  - scientific computing: multiprocessor, faster processors

5000 Bases per Page

CACACTTGCATGTGAGAGCTTCTAATATCTAAATTAATGTTGAATCATTATTCAGAAACAGAGAGCTAACTGTTATCCCATCCTGACTTTATTCTTTATG AGAAAAATACAGTGATTCC
AAGTTACCAAGTTAGTGCTGCTTGCTTTATAAATGAAGTAATATTTTAAAAGTTGTGCATAAGTTAAAATTCAGAAATAAAACTTCATCCTAAAACTCTGTGTGTTGCTTTAAATAATC
AGAGCATCTGC TACTTAATTTTTTGTGTGTGGGTGCACAATAGATGTTTAATGAGATCCTGTCATCTGTCTGCTTTTTTATTGTAAAACAGGAGGGGTTTTAATACTGGAGGAACAA
CTGATGTACCTCTGAAAAGAGA AGAGATTAGTTATTAATTGAATTGAGGGTTGTCTTGTCTTAGTAGCTTTTATTCTCTAGGTACTATTTGATTATGATTGTGAAAATAGAATTTATCC
CTCATTAAATGTAAAATCAACAGGAGAATAGCAAAAACTTATGAGATAGATGAACGTTGTGTGAGTGGCATGGTTTAATTTGTTTGGAAGAAGCACTTGCCCCAGAAGATACACAAT
GAAATTCATGTTATTGAGTAGAGTAGTAATACAGTGTGTTCCCTTGTGAAGTTCATAACCAAGAATTTTAGTAGTGGATAGGTAGGCTGAATAACTGACTTCCTATC ATTTTCAGGTT
CTGCGTTTGATTTTTTTTACATATTAATTTCTTTGATCCACATTAAGCTCAGTTATGTATTTCCATTTTATAAATGAAAAAAAATAGGCACTTGCAAATGTCAGATCACTTGCCTGTGGT
CATTCGGGTAGAGATTTGTGGAGCTAAGTTGGTCTTAATCAAATGTCAAGCTTTTTTTTTTCTTATAAAATATAGGTTTTAATATGAGTTTTAAAATAAAATTAATTAGAAAAAGGCAA
ATTACTCAATATATATAAGGTATTGCATTTGTAATAGGTAGGTATTTCATTTTCTAGTTATGGTGGGATATTATTCAGACTATAATTCCCAATGAAAAAACTTTAAAAAATGCTAGTGA
TTGCACACTTAAAACACCTTTTAAAAAGCATTGAGAGCTTATAAAATTTTAATGAGTGATAAAACCAAATTTGAAGAGAAAAGAAGAACCCAGAGAGGTAAGGATATAACCTTACC
AGTTGCAATTTGCCGATCTCTACAAATATTAATATTTATTTTGACAGTTTCAGGGTGAATGAGAAAGAAACCAAAACCCAAGACTAGCATATGTTGTCTTCTTAAGGAGCCCTCCCCT
AAAAGATTGAGATGACCAAATCTTATACTCTCAGCATAAGGTGAACCAGACAGACCTAAAGCAGTGGTAGCTTGGATCCACTACTTGGGTTTGTGTGTGGCGTGACTCAGGTAATCT
CAAGAATTGAACATTTTTTTAAGGTGGTCCTACTCATACACTGCCCAGGTATTAGGGAGAAGCAAATCTGAATGCTTTATAAAAATACCCTAAAGCTAAATCTTACAATATTCTCAAG
AACACAGTGAA ACAAGGCAAAATAAGTTAAAATCAACAAAAACAACATGAAACATAATTAGACACACAAAGACTTCAAACATTGGAAAATACCAGAGAAAGATAATAAATAT
TTTACTCTTTAAAAATTTAGTTAAAAGCTTAAACTAATTGTAGAGAAAA AACTATGTTAGTATTATATTGTAGATGAAATAAGCAAAACATTTAAAATACAAATGTGATTACTTAAAT
TAAATATAATAGATAATTTACCACCAGATTAGATACCATTGAAGGAATAATTAATATACTGAAATACAGGTCAGTAGAATTTTTTTCAATTCAGCATGGAGATGTAAAAAATGAAAA
TTAATGCAAAAAATAAGGGCACAAAAAGAAATGAGTAATTTTGATCAGAAATGTATTAAAATTAATAAACTGGAAATTTGACATTTAAAAAAAGCATTGTCATCCAAGTAGATGTG
TCTATTAAATAGTTGTTCTCATATCCAGTAATGTAATTATTATTCCCTCTCATGCAGTTCAGATTCTGGGGTAATCTTTAGACATCAGTTTTGTCTTTTATATTATTTATTCTGTTTACTAC
ATTTTATTTTGCTAATGATATTTTTAATTTCTGACATTCTGGAGTATTGCTTGTAAAAGGTATTTTTAAAAATACTTTATGGTTATTTTTGTGATTCCTATTCCTCTATGGACACCAAGGCT
ATTGACATTTTCTTTGGTTTCTTCTGTTACTTCTATTTTCTTAGTGTTTATATCATTTCATAGATAGGATATTCTTTATTTTTTATTTTTATTTAAATATTTGGTGATTCTTGGTTTTCTCAGCC
ATCTATTGTCAAGTGTTCTTATTAAGCATTATTATTAAATAAAGATTATTTCCTCTAATCACATGAGAATCTTTATTTCCCCCAAGTAATTGAAAATTGCAATGCCATGCTGCCATGTGG
TACAGCATGGGTTTGGGCTTGCTTTCTTCTTTTTTTTTTAACTTTTATTTTAGGTTTGGGAGTACCTGTGAAAGTTTGTTATATAGGTAAACTCGTGTCACCAGGGTTTGTTGTACAGATCA
TTTTGTCACCTAGGTACCAAGTACTCAACAATTATTTTTCCTGCTCCTCTGTCTCCTGTCACCCTCCACTCTCAAGTAGACTCCGGTGTCTGCTGTTCCATTCTTTGTGTCCATGTGTTCTC
ATAATTTAGTTCCCCACTTGTAAGTGAGAACATGCAGTATTTTCTAGTATTTGGTTTTTTGTTCCTGTGTTAATTTGCCCAGTATAATAGCCTCCAGCTCCATCCATGTTACTGCAAAGAA
CATGATCTCATTCTTTTTTATAGCTCCATGGTGTCTATATACCACATTTTCTTTATCTAAACTCTTATTGATGAGCATTGAGGTGGATTCTATGTCTTTGCTATTGTGCATATTGCTGCAAG
AACATTTGTGTGCATGTGTCTTTATGGTAGAATGATATATTTTCTTCTGGGTATATATGCAGTAATGCGATTGCTGGTTGGAATGGTAGTTCTGCTTTTATCTCTTTGAGGAATTGCCATG
CTGCTTTCCACAATAGTTGAACTAACTTACACTCCCACTAACAGTGTGTAAGTGTTTCCTTTTCTCCACAACCTGCCAGCATCTGTTATTTTTTTGACATTTTAATAGTAGCCATTTTAACT
GGTATGAAATTATATTTCATTGTGGTTTTAATTTGCATTTCTCTAATGATCAGTGATATTGAGTTTGTTTTTTTTCACATGCTTGTTGGCTGCATGTATGTCTTCTTTTAAAAAGTGTCTGTT
CATGTACTTTGCCCACATTTTAATGGGGTTGTTTTTCTCTTGTAAATTTGTTTAAATTCCTTATAGGTGCTGGATTTTAGACATTTGTCAGACGCATAGTTTGCAAATAGTTTCTCCCATTC
TGTAGGTTGTCTGTTTATTTTGTTAATAGTTTCTTTTGCTATGCAGAAGCTCTTAATAAGTTTAATGAGATCCTGATATGTTAGGCTTTGTGTCCCCACCCAAATCTCATCTTGAATTATA
TCTCCATAATCACCACATGGAGAGACCAGGTGGAGGTAATTGAATCTGGGGGTGGTTTCACCCATGCTGTTCTTGTGATAGTGAATGAGTTCTCACGAGATCTAATGGTTTTATGAGG
GGCTCTTCCCAGCTTTGCCTGGTACTTCTCCTTCCTGCCGCTTTGTGAAAAAGGTGCATTGCGTCCCTTTCACCTTCTTCTATAATTGTAAGTTTCCTGAGGCCTTCCCAGCCATGCTGAA
CTTCAAGTCAATTAAACCTTTTTCTTTATAAATTACTCAGTCTCTGGTGGTTCTTTATAGCAGTGTGAAAATGGACTAATGAAGTTCCCATTTATGAATTTTTGCTTTTGTTGCAATTGCTT
TTGACATCTTAGTCATGAAATCCTTGCCTGTTCTAAGTACAGGACGGTATTGCCTAGGTTGTCTTCCAGGGTTTTTCTAATTTTGTGTTTTGCATTTAAGTGTTTAATCCATCTTGAGTTGA
TTTTTGTATATTGTGTAAGGAAGGGGTCCAGTTTCAATCTTTTGCATATGGCTAGTTAGTTATCCCAGTACCATTTATTGAAAAGACAGTCTTTTCCCCATCGCTCGTTTTTGTCAGTTTT
ATTGATGATCAGATAATCATAGCTGTGTGGCTTTATTTCTGGGTTCTTTATTCTGTTCTATTGGTTTATGTCCCTGTTTTTGTGCCAGTACCATGCTGTTTTGGTTAACATAGCCCTGTAGT
ATAGTTTGAGGTCAGATAGCCTGATGCTTCCAGCTTTGTTCTTTTTCTTAAGATTGCCTTGGCTATTTGGCCTCTTTTTTGGTTCCACATGAATTTTAAAACAGTTGTTTCTAGTTTTTGAA
GAATGTCATTGGTAGTTTGATAGAAATAGCATTTAATCTGTAAATTGATTTGTGCAGTATGGCCTTTTAATGATATTGATTCTTCCTATCCATGAGCATGATATGTTTTCCATTTTGTTTG
TATCCTCTCTGATTTCTTTGTGCAGTGTTTTGTAATTCTCAT TGTAGAGATTTTTCACCTCCCTGGTTAGTTGTATTTTACCCTAGATATTT TATTCTTTTTGTGAAAATTGTGAATGGGAT
TGCCTTCCTGATTTGACTGC CAGCTTGGTTACTGTTGGTTTATAGAAATGCTAGTGATTTTTGTACATTG ATTTTCTTTCTAAAACTTTGCTGAAGTTTTTTTTATTAGCAGAAGGAGCT
TTGGGGCTGAGACTATGGGGTTTTCTAGATATAGAATCATGTCAGCTTCAAATAGGGATAATTTTACTTCCTCTCTTCCTATTTGGATGCCCTTTATTTCTTTCTCTTGCCTGATTACTCTG
GCTGGGATTTCCTATGTTGAATAGGAGT CATGAGAGAGGGCATCAAATCTACACATATCAAATACTAACCTTGAATGTCTAGATATTT TATTCTTTTTGTGAAAATTGTGAATGGGAT

The magnitude of genomic data in an analysis!

- 3 pallets with 40 boxes per pallet x 5000 pages per box x 5000 bases per page = 3,000,000,000 bases!

- To get accurate sequence
- requires 6-fold coverage

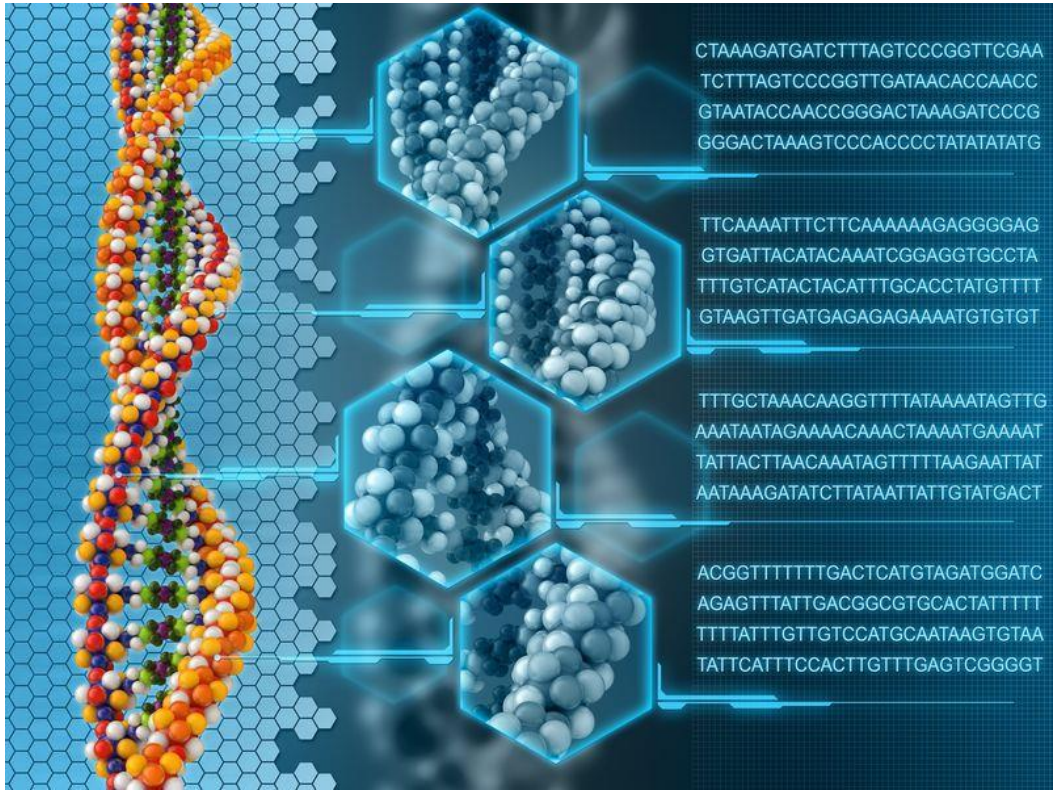
- Now: Shred 18 pallets and reassemble
- Really need Bioinformatics

Need for Bioinformaticians?

- Many of the processes we use today involve bioinformatics
- Is it necessary to have a bioinformatics specialist in the laboratory?
- Will the current software or near future software be sufficient to stand alone?
- Is it necessary to increase education/training in aspects of bioinformatics?
- What features should be added?
- Need for community to start using!

Bioinformatics

- Allele calls
- Alignment
- Strand bias
- Coverage
- Noise
- Thresholds
- ….



Objectives

- Sample to answer
  - Simplicity
  - Flexibility
  - Data and information management
- What do the data mean?

STR Allele Calling Software

- Compare back to nominal allele length
- Needed better approaches
- More versatile, more flexible

Forensic Science International: Genetics 7 (2013) 409–417

Contents lists available at SciVerse ScienceDirect

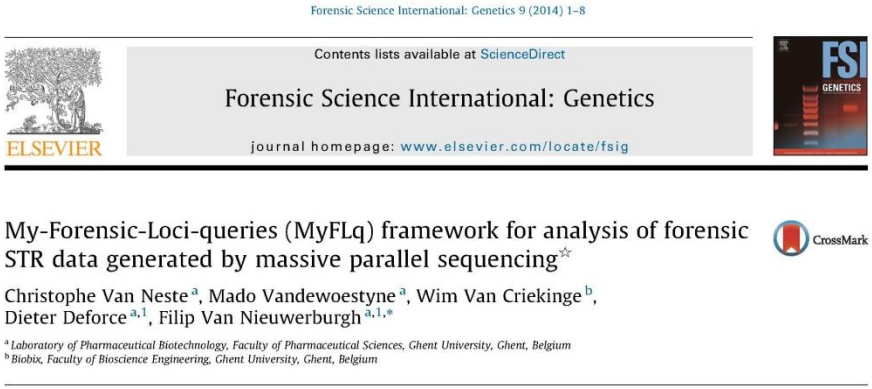
Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

David H. Warshauer[a], David Lin[b], Kumar Hari[b], Ravi Jain[b], Carey Davis[a], Bobby LaRue[a], Jonathan L. King[a], Bruce Budowle[a,c,*]

[a] Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA
[b] cBio, Inc., 37869 Abraham Street, Fremont, CA 94536, USA
[c] Center of Excellence in Genomic Medicine (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia



Forensic Science International: Genetics 9 (2014) 1–8

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

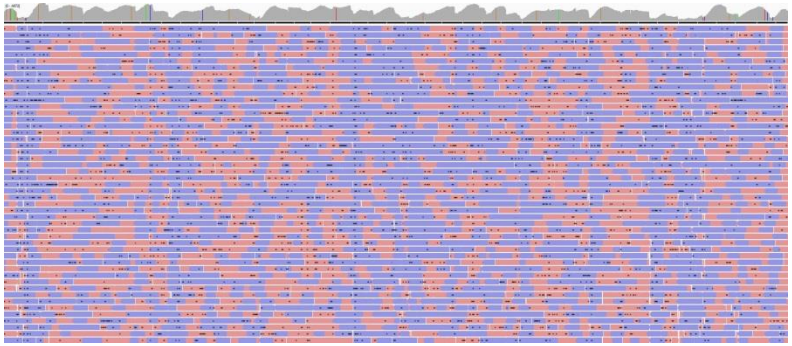
My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing☆

Christophe Van Neste[a], Mado Vandewoestyne[a], Wim Van Criekinge[b], Dieter Deforce[a,1], Filip Van Nieuwerburgh[a,1,*]

[a] Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
[b] Biobix, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

STRait Razor	2 flanks	agrep	short repeat match
MyFLq	2 primers	flexible flanking with k-mers	Modified N-W

mitoSAVE



haplogrep

Forensic Science International: Genetics 12 (2014) 122–125

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

mitoSAVE: Mitochondrial sequence analysis of variants in Excel

Jonathan L. King[a,*], Antti Sajantila[a,b], Bruce Budowle[a,c]

[a] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA
[b] Department of Forensic Medicine, Hjelt Institute, P.O. Box 40, 00014 University of Helsinki, Helsinki, Finland
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

Truth About Science Careers

ACKNOWLEDGMENTS

- Thermo Fisher
- Illumina, Inc.
- Promega Corporation
- Walter Parson

Research Team

- South Dakota Historical Society Archaeological Research Center City of Deadwood

UNT | HEALTH SCIENCE CENTER

INSTITUTE OF APPLIED GENETICS