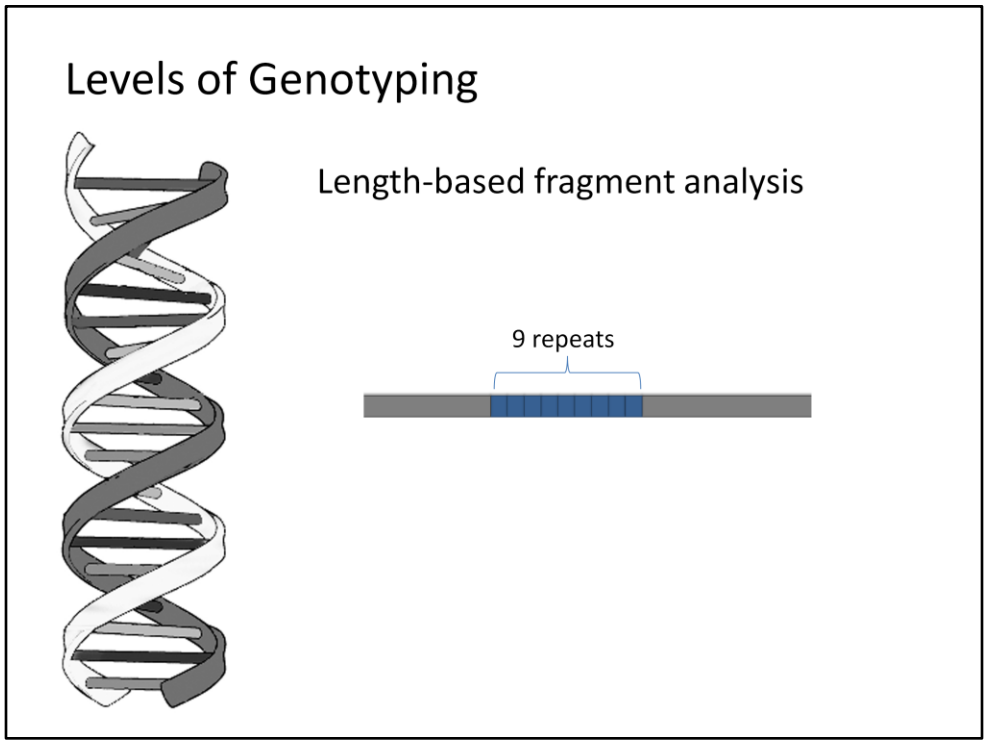
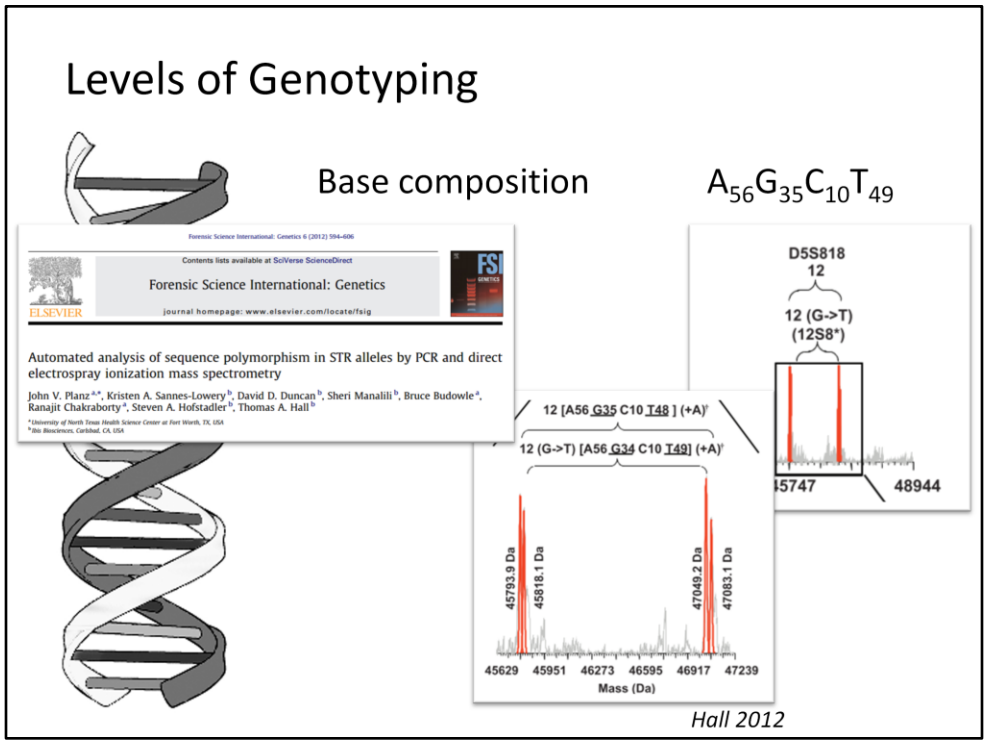
# Forensic Applications of Next Generation Sequencing


SWGDAM Meeting, January 2013
Dr. Katherine Butler Gettings
Applied Genetics Group
NIST



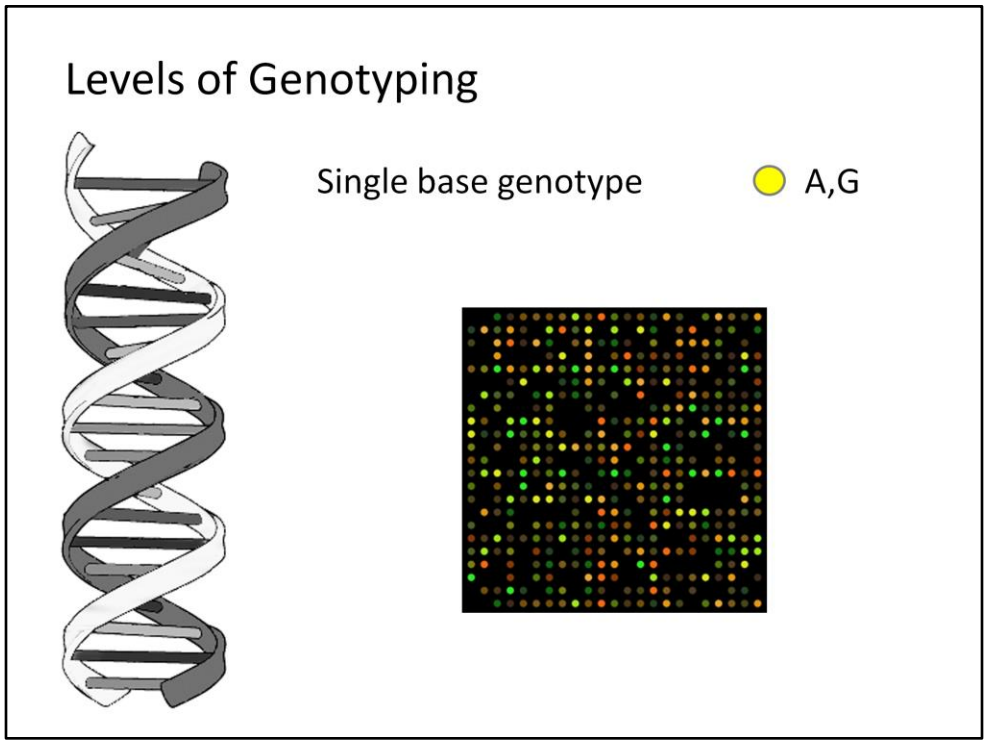

There are several different levels of genotyping that we will cover in this lecture, so this is a quick overview of these levels and the information obtained by each.

Current STR technology returns a number representing the length of a fragment.




A mass spec method that determines the mass of each fragment and genotypes by comparison to a reference.

At D5S818 we see two peaks for a 12 allele, which is expected because the forward and reverse strands have different masses due to the complimentary base compositions. But zooming in closer, we see each 12 is a doublet, and this is due to a SNP present in one copy of the individual's chromosomes and not in the other. So these two different versions of a 12 allele migrate slightly differently. We can't be sure where the SNP is but we know it exists.

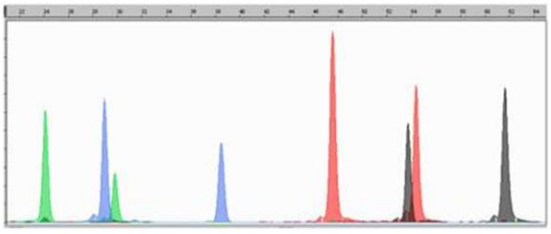



Array methods interrogate individual SNPs.

## Levels of Genotyping

Single base genotype

A,G



Snapshot methods also interrogate individual SNPs.

## Levels of Genotyping

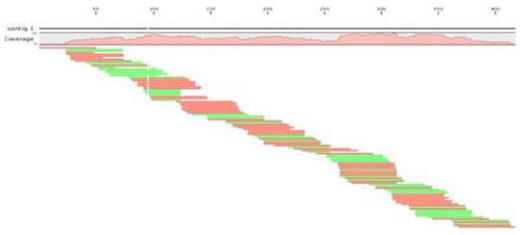
Sequence comparison to reference

263A>G


Sequence data can be compared to a reference, as in traditional Sanger sequencing of mtDNA shown here, where ideally both F and R strands are sequenced, and aligned/compared to a reference sequence.

## Levels of Genotyping

*De novo* Sequencing



De novo sequencing is how unknown genomes are sequenced, when no reference genome exists. Sequences are aligned based on their overlapping regions, and a consensus sequence is determined.

## Which genotyping methods could be employed on NGS data?

A. Length-based binning

B. Base counts

C. Single base genotype

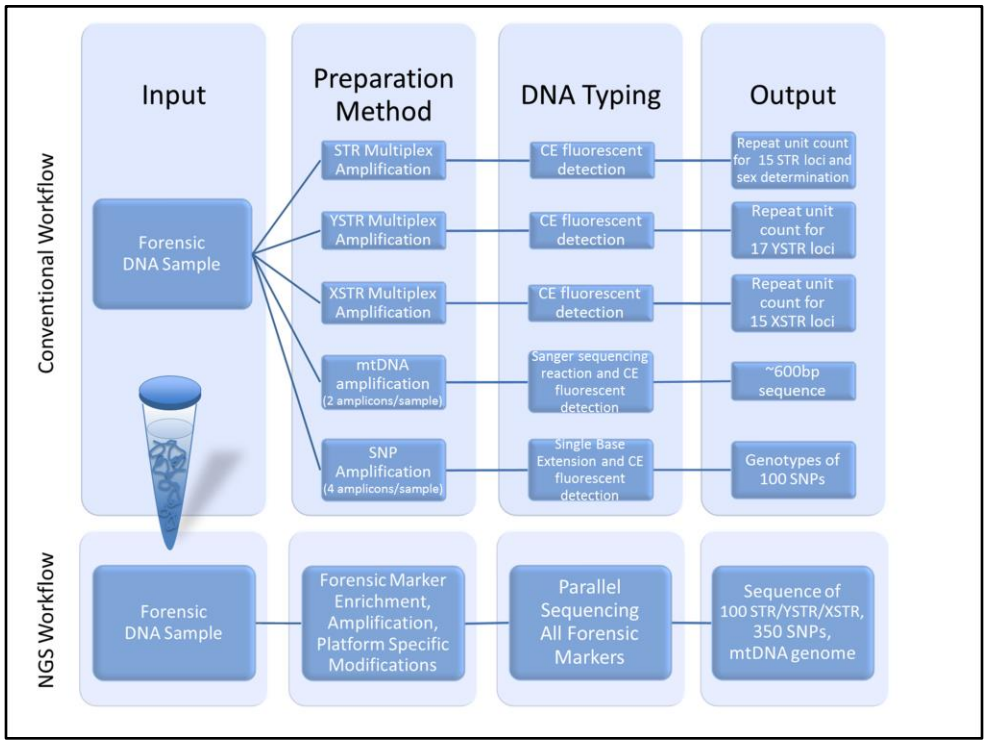
D. Sequencing comparison to a reference

E. De novo sequencing

# Use of NGS for forensic applications

Highly-parallel/high-throughput direct sequencing of forensically relevant targets

- Whole mitochondrial genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Forensically relevant SNPs
  - newer human identity applications
  - biogeographical ancestry, externally visible traits, complex kinship
  - degraded samples, mixtures, low template?
- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'




Comparison of conventional (current) forensic DNA workflows and a possible (future) NGS workflow.

# NGS Forensic Applications

Challenges

- Sample input requirements
- Library preparation methods
- Read lengths
- Data analysis
  - Assembly
  - Interpretation
  - Storage
- **Cost** and **time** per unit of information
- Privacy, disease related markers
- Validation, court acceptance

read length– primarily applies to repeat sequences

interpretation-- nomenclature

Assembly— errors, platform & bioinformatics based biases, barcoding– all need extensive validation

Validation– this is a rapidly changing technology, forensic validation would require choosing a platform

Court admissibility

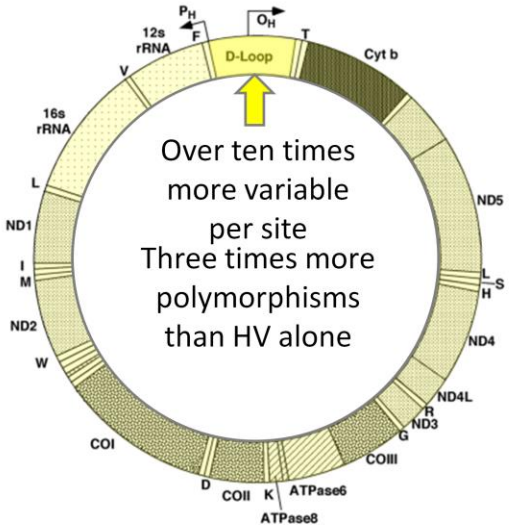
# Use of NGS for Forensic Applications

Highly-parallel/high-throughput direct sequencing of forensically relevant targets

- Whole mitochondrial genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Forensically relevant SNPs
  - newer human identity applications
  - biogeographical ancestry, externally visible traits, complex kinship
  - degraded samples, mixtures, low template?
- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'

## mtDNA Information

- Increase in variants by whole genome analysis




*(based on analysis of 3 SRM samples)*

*http://www.sas.upenn.edu/~tgschurr/labwork/labwork_text.html*

3X more polymorphisms does not mean 3X the discriminating power, but some of these additional polymorphisms may help resolve common haplotypes, or may provide the two polymorphisms needed to exclude an individual.

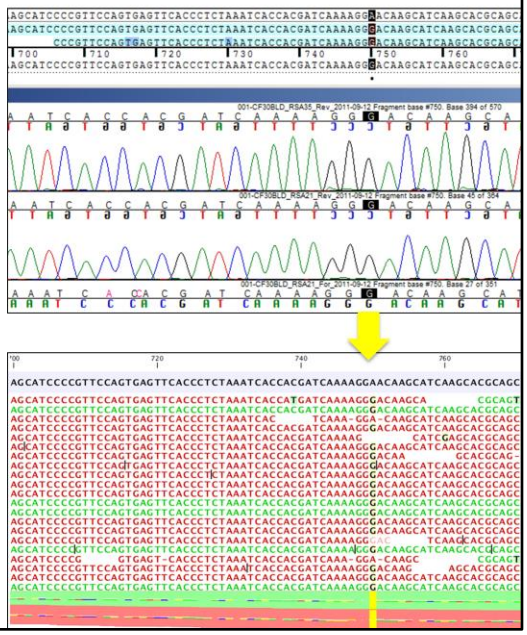
## mtDNA Information

Current Method

- Sequence based on chromatogram
- Consensus of one forward and one reverse

NGS

- Sequence based on thousands of individual reads
- Improved sensitivity:
  - Mixture detection
  - Low level heteroplasmy



A “normal” variant site with 100% frequency.

Sanger sequencing results in a chromatogram which shows total signal for that sequencing sample. The sample genotype is a consensus of generally one F & one R.

NGS data is like thousands of sequencing reactions overlapping at each site. Because NGS looks at base calls rather than fluorescence, we can “see” and quantify rare variants, potentially improving sensitivity. What court issues will this raise?

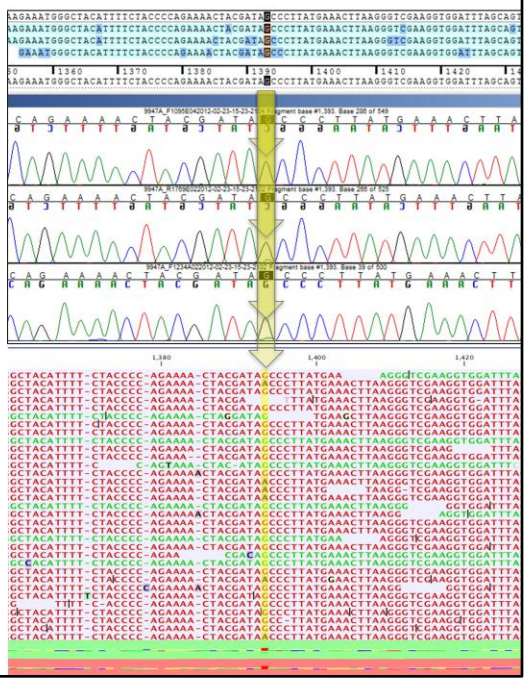
## mtDNA Information

Current Method

- Minor peaks may not be reproducible
- SRM 2392 9947a, 1393 G/A heteroplasmy

NGS

- More consistent detection of minor genotypes
- Validation important
  - Variant calling thresholds
  - Characterizing noise



A low level heteroplasmy.

Sanger data will sometimes show a minor type, sometimes not. Makes it difficult to interpret heteroplasmies and mixtures.

NGS will more consistently detect minor types, even very low level ~1%, but must be able to distinguish from noise. Validation of variant calling thresholds will be important.

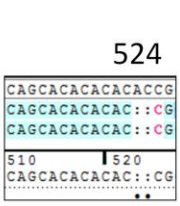
## mtDNA Challenges

- Known problem areas

C-stretches




524



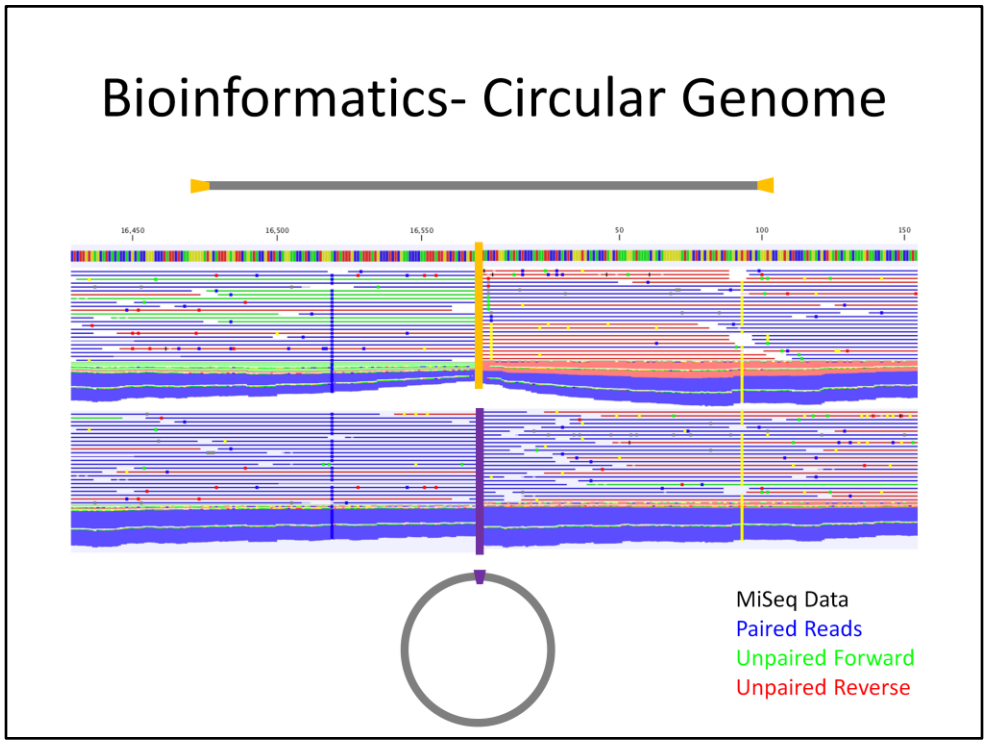

- Nomenclature
  - reference sequence?
  - reporting whole genome
- Using haplogroup to QC
- Circular genome


Problem areas for Sanger and NGS. The NGS bioinformatics don't use "forensic" conventions such as shifting indels to the 5' end of homopolymers or repeat regions. These examples show that insertions are placed at the 3' end. This would cause an apparent discrepancy between Sanger and NGS data for the same sample, but it is a matter of adapting the NGS bioinformatics to forensic conventions.

Difficulty "sequencing" indels is due to use of reference genome, because they only exist when assembled to a fixed genome.

Nomenclature– will we continue the convention of reporting differences to rCRS or simply report the whole genome sequence and do direct comparisons of samples?

With WGS of mtDNA, we'll be looking at polymorphisms in regions that are less familiar than HV regions, so could benefit from haplogroup based QC analysis. Resources such as EMPOP (EDNAP mt pop db, EDNAP=European DNA Profiling group) could help by using phylogenetic analysis of observed haplotypes to pinpoint possible technical errors. Currently EMPOP only allows evaluation of HV region haplotypes.




One of the bioinformatics packages we are using allows for setting the reference genome as a circular molecule. This is the same sample analyzed to a linear rCRS (top) versus a circularized rCRS (bottom). Notice the improved coverage across what we designate the “end” and “beginning” of the genome when the reference is circularized.

## NGS Forensic Applications

Whole Genome mtDNA sequencing

- ☑ Sample type amenable to library preparation?
- ☑ Sample type amenable to sequencing platforms?
- ☑ Sample type amenable to bioinformatics?
- ☑ Improvement over current method?

Amenability to library preparation– degraded samples may not produce results for entire genome, but in those cases, a HV region amplification and sequencing approach could be used with NGS technology.

Amenability to bioinformatics– yes with tweaking and making the NGS bioinformatics consistent with forensic conventions.

Improvement over current method– yes, NGS will allow for multiplexing of many samples, and facilitate whole genome sequencing, which is very labor intensive with Sanger technology.

# Use of NGS for forensic applications

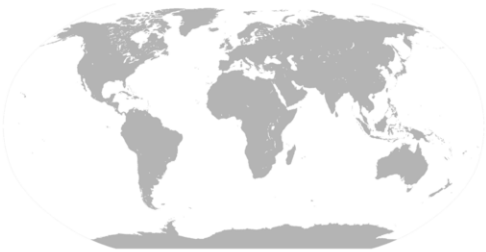
Highly-parallel/high-throughput direct sequencing of forensically relevant targets

- Whole mitochondrial genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Forensically relevant SNPs
  - newer human identity applications
  - biogeographical ancestry, externally visible traits, complex kinship
  - degraded samples, mixtures, low template?
- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'

Snps are new, complimentary information to our current forensic markers.

## SNP Information

- IISNP-Individual
- AISNP-Ancestry
- LISNP-Lineage
- PISNP-Phenotype



Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics Supplement Series 1 (2008) 471–472
www.elsevier.com/locate/FSIGSS
Research article

Report on ISFG SNP Panel Discussion

J.M. Butler[a,*], B. Budowle[b], P. Gill[c], K.K. Kidd[d], C. Phillips[e], P.M. Schneider[f], P.M. Vallone[a], N. Morling[g]

[a] National Institute of Standards and Technology, Gaithersburg, MD, USA
[b] FBI Laboratory, Quantico, VA, USA
[c] Forensic Science Service, Birmingham, UK
[d] Yale University, New Haven, CT, USA
[e] University of Santiago de Compostela, Spain
[f] University of Cologne, Germany
[g] University of Copenhagen, Denmark

Categories first described by Kidd 2007, definitions found in Report on ISFG SNP Panel Discussion. IISNPs- Polymorphisms that collectively have very low probability of two individuals having the same multi-locus genotype (except for identical twins). Redundant to STRs but no core loci.

AISNPs- collectively can give a high probability of an individual's ancestry being from one part of the world. Good for investigative lead– resolution mainly at the continental level currently.

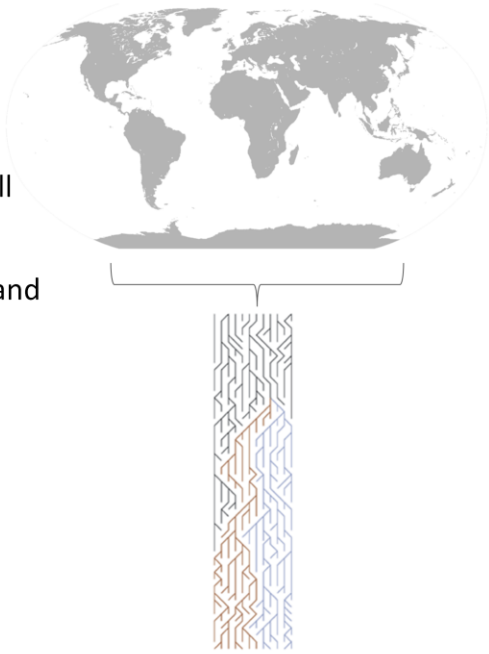
LISNPs- sets of tightly linked snps that function as multiallelic markers that can identify relatives with higher probabilities than biallelic snps

PISNPs- provide a high probability that the individual has particular phenotypes, such as a particular skin color, hair color, eye color, etc.

AI & PISNPs- investigative leads to prioritize suspect processing, corroborate witness testimony, determine relevance of evidence

## SNP Information

- Individual Identification
  - Balancing has occurred in all populations
  - Low F statistics within ($F_{IS}$) and among ($F_{ST}$) populations
  - High heterozygosity



F statistics measure population differentiation, estimated by genetic data—
F IS measures the variance in allele frequencies among individuals compared to average variance in their subpopulation
F ST measures the variance in allele frequencies among subpopulations compared to the average variance in the total population

Ideal IISNPs are low Fis & Fst (zero), and high heterozygosity (highest possible =0.5, eg AA=0.25, AG=0.5, GG=0.25)

## SNP Information

- Individual Identification

  *Pakstis 2010, Kidd 2012*

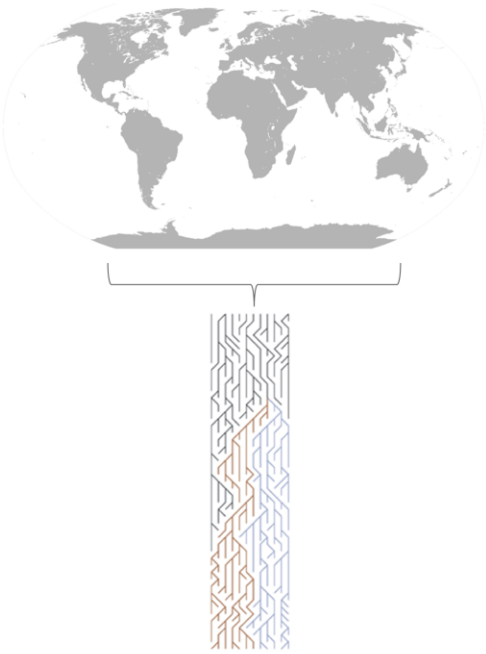
  – 45 unlinked SNPs

  – $F_{ST}$ below ~0.07

  – Avg het >0.4

  – RMP $10^{-15}$ to $10^{-18}$

  *in 44 populations*



RMP is on the level of CODIS STR loci, and SNPs have a benefit of lower mutation rate, which can be helpful in kinship analysis.

These references are NIJ funded work that has been ongoing for past decade, but haven't had a good way to genotype these SNPs. Now all that work begins to pan out when we can quickly genotype a large number of SNPs.

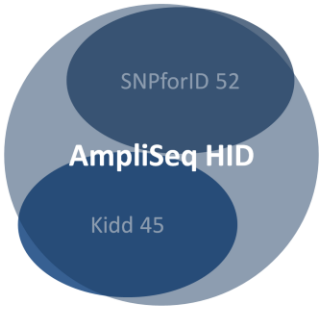
## SNP Information

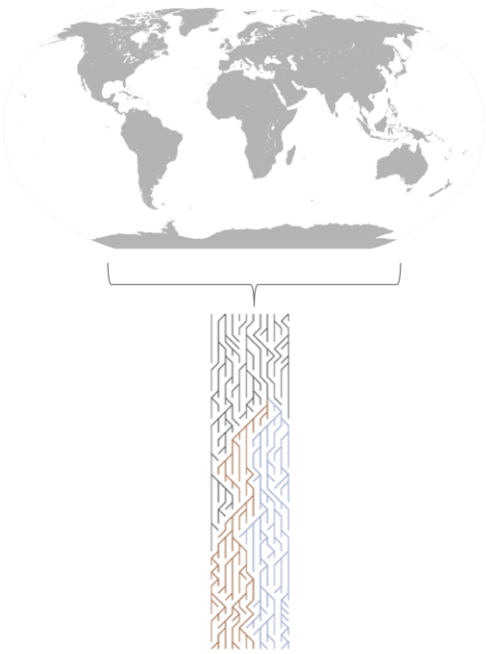
- Individual Identification

  *Ion AmpliSeq HID kit (PGM) v 2.3*

  – 90 autosomal SNPs

  – 30 Y-chromosome SNPs



48 of SNPforID 52
37 of Kidd's 45
5 not in either panel

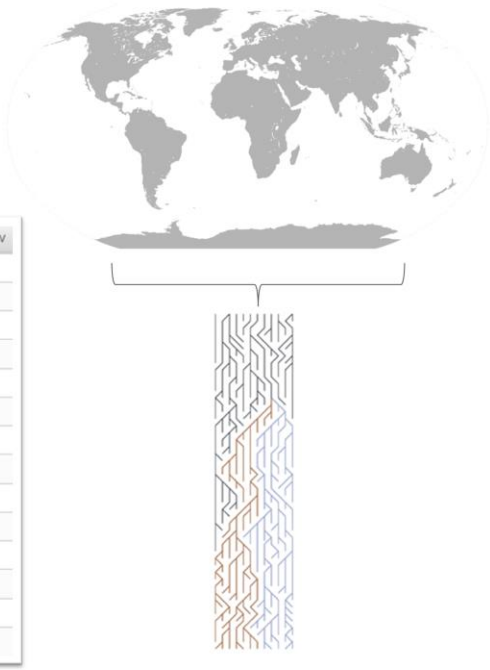
NIST staff recently ran a pre-release version of this kit on our SRM samples (N=14) and obtained around 1000-2000X coverage per sample per locus.

## SNP Information

- Individual Identification

  *Ion AmpliSeq HID kit (PGM) v 2.3*

| Gen... | Ref | Variant | Var Fr... | Qual | Cov | Ref Cov | Var Cov |
|---|---|---|---|---|---|---|---|
| T/T | A | C,G,T | 100.0% | 100 | 395 | 0 | 395 |
| G/A | G | A,C,T | 46.6% | 100 | 399 | 213 | 186 |
| T/C | T | A,C,G | 50.3% | 100 | 400 | 199 | 201 |
| T/T | T | A,C,G | 0.0% | 100 | 399 | 399 | 0 |
| T/G | T | A,C,G | 50.3% | 100 | 398 | 198 | 200 |
| T/T | T | A,C,G | 0.0% | 100 | 400 | 400 | 0 |
| C/G | C | A,G,T | 47.8% | 100 | 400 | 209 | 191 |
| T/A | T | A,C,G | 46.6% | 100 | 399 | 213 | 186 |
| G/A | G | A,C,T | 50.8% | 100 | 400 | 197 | 203 |
| A/A | G | A,C,T | 100.0% | 100 | 399 | 0 | 399 |
| A/A | A | C,G,T | 0.0% | 100 | 400 | 400 | 0 |
| G/A | G | A,C,T | 50.1% | 100 | 399 | 199 | 200 |
| A/A | A | C,G,T | 0.0% | 100 | 398 | 398 | 0 |
| G/A | G | A,C,T | 49.0% | 100 | 400 | 204 | 196 |

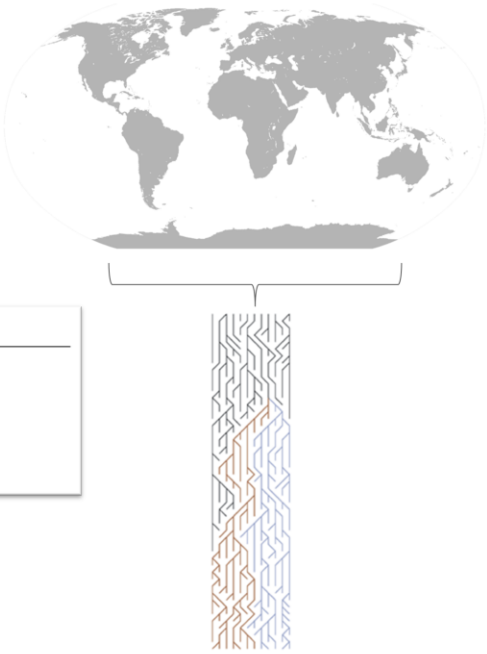


There will be an analysis plug-in for the PGM server that returns the SNP genotypes. This is a view of some of the autosomal SNP data we recently generated (rs numbers not visible here but are present in actual report). First row is a homozygous SNP, differing from the reference. Heterozygotes appear balanced (near 50%).

## SNP Information

- Individual Identification

  *Ion AmpliSeq HID kit*



Int J Legal Med (2013) 127:1079–1086
DOI 10.1007/s00414-013-0879-7

ORIGINAL ARTICLE

**Single nucleotide polymorphism typing with massively parallel sequencing for human identification**

Seung Bum Seo · Jonathan L. King · David H. Warshauer · Carey P. Davis · Jianye Ge · Bruce Budowle

This article from this past year presents results from an earlier version of this kit.

## SNP Information

- Individual Identification

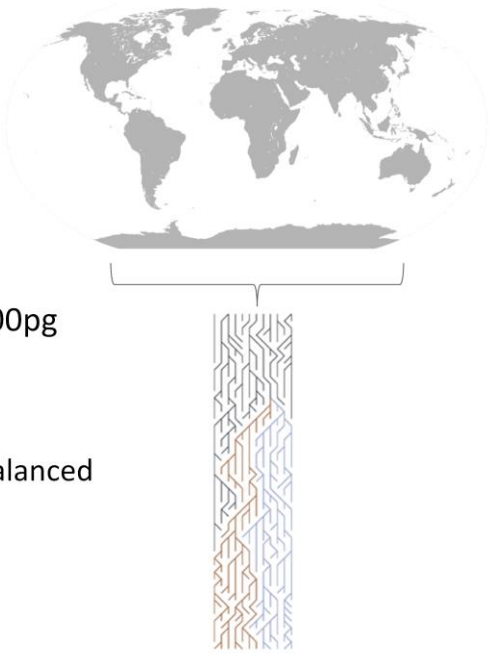
  *Ion AmpliSeq HID kit*

  *Seo 2013*

  – 4 samples: 10ng, 1ng, 100pg

  – 10ng & 1ng
    - all present
    - 1 sample at 1 locus imbalanced

  – 100pg
    - 99% present
    - 97% balanced



The authors show complete results down to 1ng and close to complete results even at 100pg of input DNA.

In terms of molecular evolution, what selective forces have IISNPs undergone?

A. Positive Selection

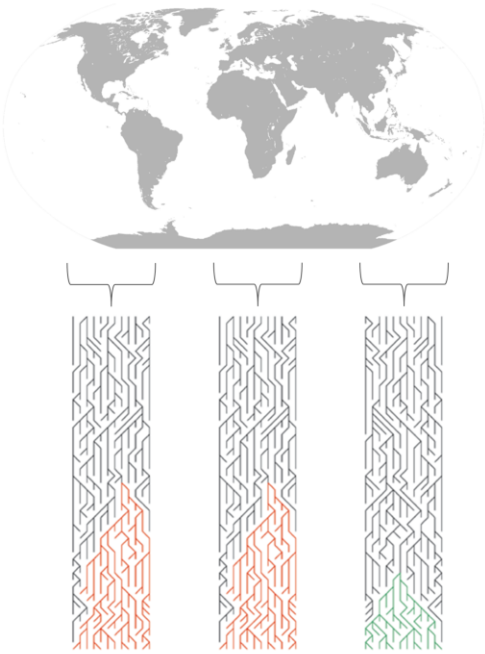
B. Neutral (No Selection)

C. Negative Selection

D. Reverse Selection

## SNP Information

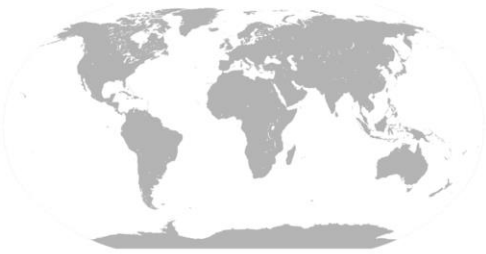
- Ancestry Information
  - Population specific fixation has occurred
  - Low $F_{IS}$
  - High $F_{ST}$
  - Low heterozygosity

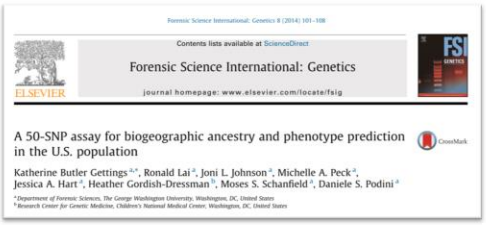


Fixation may be the result of negative “purifying” selection, eg malaria resistance snps in subsaharan africa

## SNP Information

- Ancestry Information





Forensic Science International: Genetics 8 (2014) 101–108

Contents lists available at ScienceDirect

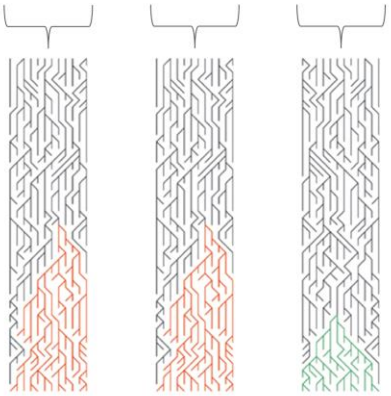
Forensic Science International: Genetics

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

A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population

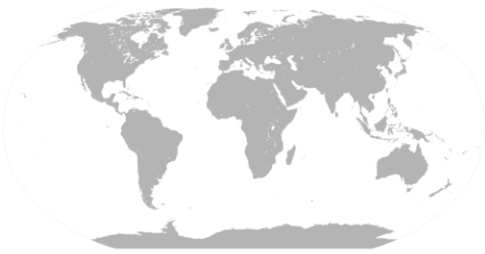
Katherine Butler Gettings [a,*], Ronald Lai [a], Joni L. Johnson [a], Michelle A. Peck [a], Jessica A. Hart [a], Heather Gordish-Dressman [b], Moses S. Schanfield [a], Daniele S. Podini [a]

[a] Department of Forensic Sciences, The George Washington University, Washington, DC, United States
[b] Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, United States



An article showing a panel of SNPs designed for BGA in the US population.

## SNP Information



- Ancestry Information

  *Gettings 2014*

  - 32 SNP ancestry model
    - European
    - African American
    - Hispanic
    - East Asian
  - Highest overall $F_{ST}$=0.774
  - Highest pairwise $F_{ST}$=0.886

The model is designed to predict the four primary populations of interest in the US.
Highest average Fst in alfred is 0.774 for rs2714758
Highest pairwise Fst from our training set is 0.886 for rs1426654 Asian compared to European

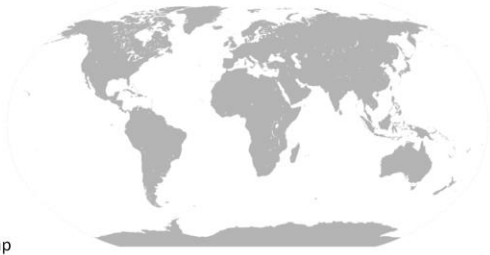
## SNP Information

- Ancestry Information

  *Gettings 2014*

  – RMP/LR prediction

    - http://mathgene.usc.es/snipper/index.php



**Calculation results**

**Processing Excel file: "Training Set Ancestry 32 SNP.xlsx".**

**********

**Executing the query with your custom population file and the following 32 SNPs of the individual to classi**
CCGGGGTTGAAACTAAGGAATTCCCCTTTTTTTTGGTTCTGTGAAACCAAGGCCTTGTCTTTAA

The **-log(LIKELIHOOD)** (lower is best) and PERCENTILE (percent of population samples with lower likelihoods than individual subm

| | | |
|---|---|---|
| **European** | 33.486693 | 1.13% |
| **EastAsian** | 101.202260 | 0.00% |
| **AfricanAA** | 94.343492 | 0.00% |
| **HisNatAmer** | 42.139509 | 1.28% |

This profile is 5,726.2492 times more likely European than HisNatAmer, and 269,012,209,794,635,348,794,408,960.0000 times more likely European than AfricanAA.

Therefore, the profile should be **European**.

*This profile is 5,726 times more likely European than Hispanic/ Native American, $2.7x10^{26}$ times more likely European than African American, etc.*

"Snipper" from Chris Phillips' lab-- Interpretation with RMP/LR statistic, similar to statistics currently used for STR. Allows for the input of any SNP data set by which to compare an unknown sample, or user can choose to import a custom data set. A supplement to Gettings 2014 contains a data set with the four previously mentioned populations.

A likelihood approach such as this could be more transparent for investigators than a % prediction.

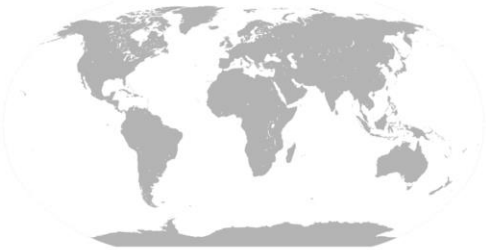
## SNP Information

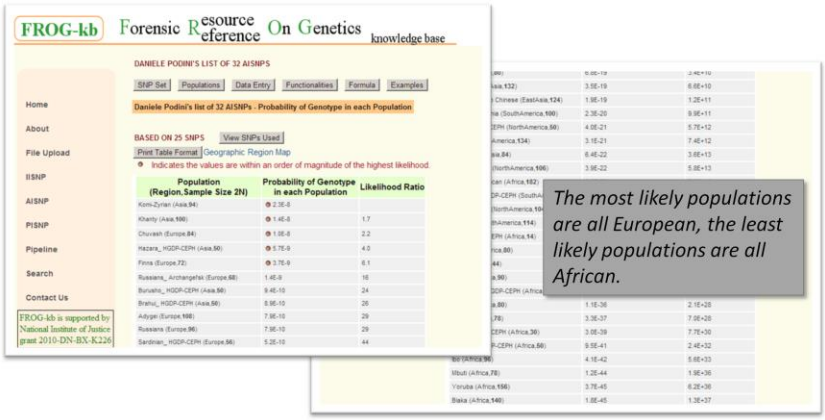
- Ancestry Information

  *Gettings 2014*

  – RMP/LR prediction

    - http://frog.med.yale.edu/FrogKB/

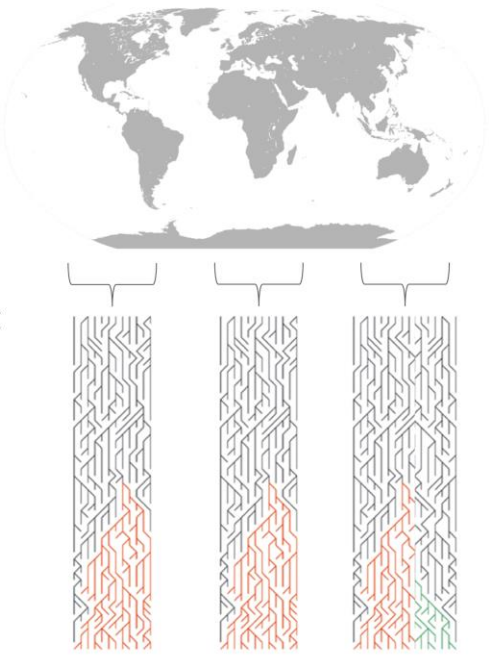





Another similar approach from the Kidd lab, using allele frequencies in 80+ world populations to determine the RMP of the unknown sample in each world population and then comparing the RMPs.

## SNP Information

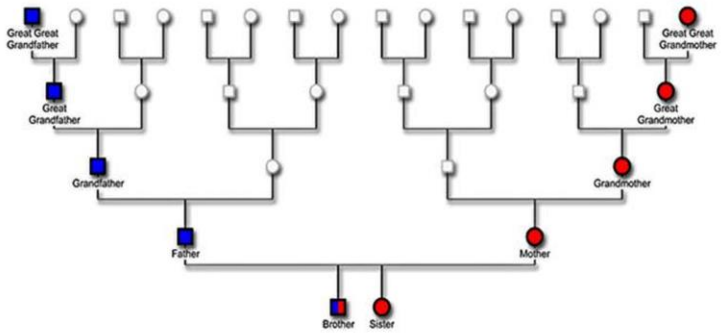
- Lineage Information
  - Indicate IBD
  - biallelic markers cannot distinguish IBS/IBD



Tightly linked snps that can give information about familial relationships because when two individuals have the same multilocus genotype, this can indicate ibd whereas a single snp (with only three possible genotypes) cannot distinguish ibs vs ibd. mt & Y snps are great example

## SNP Information

- Ancestry versus Lineage

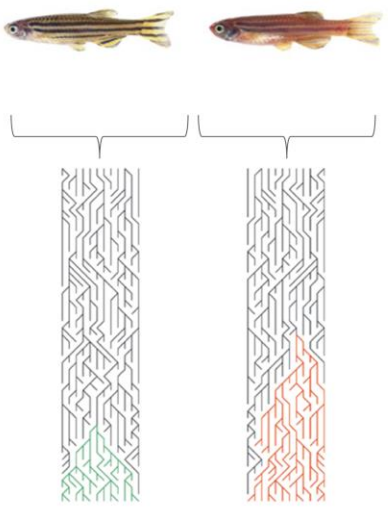



This illustrates the difference between information content for lineage markers, whether LISNPs on autosomes (such as microhaplotypes) or mt/Y data, compared to ancestry informative markers.

Using only mito and Y data, we only obtain information on 1/8$^{th}$ of a male's overall ancestry four generations prior, and for a female, mito only gives 1/16$^{th}$ of overall ancestry four generations prior.

## SNP Information

- Phenotype Information
  - Phenotype specific mutation has occurred
  - Often also ancestry informative

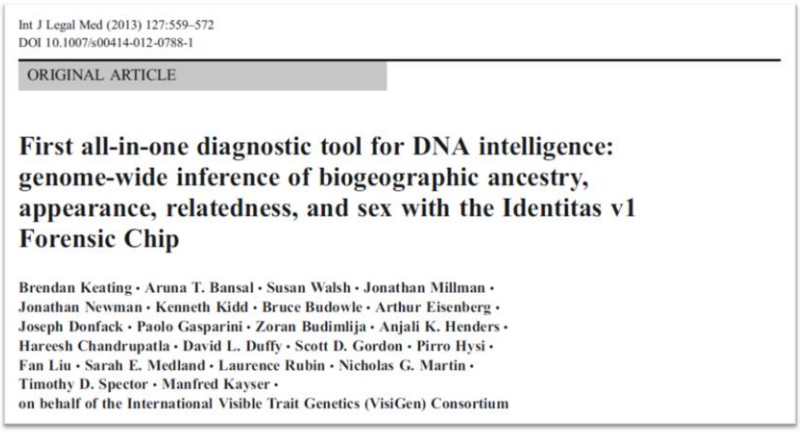


SLC24A5 gene codes for a solute carrier protein, involved in cation exchange.

Nonsynonymous mutations in this gene disrupt melanogenesis & results in the golden phenotype in zebrafish.

One polymorphism in this gene rs1426654 is fixed in the European population & appears to be a major factor in lighter skin pigmentation among Europeans– this is an example of positive (adaptive) selection.

## SNP Information



Int J Legal Med (2013) 127:559–572
DOI 10.1007/s00414-012-0788-1

ORIGINAL ARTICLE

**First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip**

Brendan Keating · Aruna T. Bansal · Susan Walsh · Jonathan Millman · Jonathan Newman · Kenneth Kidd · Bruce Budowle · Arthur Eisenberg · Joseph Donfack · Paolo Gasparini · Zoran Budimlija · Anjali K. Henders · Hareesh Chandrupatla · David L. Duffy · Scott D. Gordon · Pirro Hysi · Fan Liu · Sarah E. Medland · Laurence Rubin · Nicholas G. Martin · Timothy D. Spector · Manfred Kayser · on behalf of the International Visible Trait Genetics (VisiGen) Consortium

Array Method – Not Sequencing
(but shows promise of SNPs)

This recent article shows a SNP array designed for forensic use.

## SNP Information

*Keating 2013*

- >3000 blinded samples tested
- 201,173 SNP Chip
  - Autosomal (192,658)
  - X chromosome (5,075)
  - Y chromosome (3,012)
  - Mitochondrial (428)

This is a brute force method, using over 200K SNPs.

>90% call rate = passing platform QC

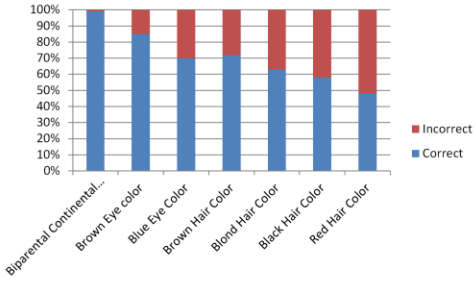
SNPs with overlapping information, and the prediction models are adaptable to the SNPs that produce data.

Sensitivity around 10ng input required, less for only BGA prediction.

## SNP Information

*Keating 2013*

- >3000 blinded samples tested
- 201,173 SNP Chip
  - Biparental Continental Ancestry
  - Eye Color & Hair Color




Graph is based on paper, not from paper. BCA was determined with PCA from 80K+ autosomal SNPs, 484 Y SNPs, and 280 mtDNA SNPs. Eye color based on Irisplex 6 SNPs, slightly lower success than the original publication, authors attribute to the samples being self-reported phenotypes. Hair color, particularly red hair is low likely due to the array missing 4 of the Hirisplex 22 SNPs, all from the MC1R gene, which is largely responsible for the red hair phenotype.

## SNP Information

| | | | |
|---|---|---|---|
| **SAMPLE QUALITY** | | | |
| **Pass** | Available genotypes | 99.7% | |
| | Laboratory quality check (QC) | Pass | Samples that pass QC are highly informative |
| **SEX** | | | |
| **Male** | X-Chromosome heterozygosity | Low | |
| | Y-Chromosome genotypes | Detected | |
| **ANCESTRY** | | | |
| | Estimated African ancestry (%) | 1% | |
| | Estimated European ancestry (%) | 95% | |
| **European** | Estimated East Asian ancestry (%) | 1% | |
| | Estimated South Asian ancestry (%) | 2% | |
| | Estimated South American ancestry (%) | 2% | |
| **HAPLOGROUPS** | | | |
| **HV0** | Mitochondrial haplogroup | HV0 | Western Eurasia |
| **R1b** | Y-Chromosome haplogroup | R1b-P310 | Europe |
| **EYE COLOR** | | | |
| **Blue/Intermediate** | Probability of brown eyes (%) | 1% | Blue or intermediate color eyes are inferred |
| | Probability of blue or intermediate color eyes (%) | 99% | |
| **HAIR COLOR** | | | |
| **Dark hair** | Probability of brown or black hair (%) | 94% | Brown or black hair is inferred |
| | Probability of blonde or red hair (%) | 6% | |

http://identitascorp.com/uploads/Sample_report_form.pdf

An example report from Identitas, showing more generalized categories to improve phenotype predictions.

# SNP Issues

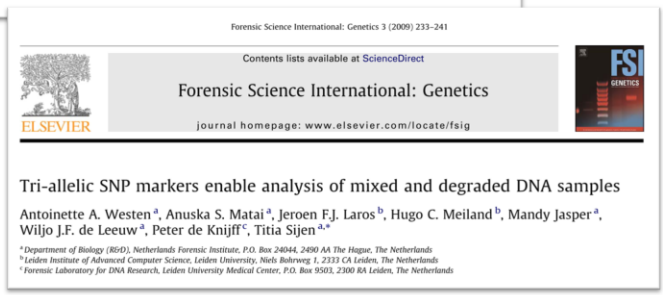
## Mixture Detection

Forensic Sci Med Pathol (2007) 3:200–205
DOI 10.1007/s12024-007-0018-1

ORIGINAL PAPER

**STRs vs. SNPs: thoughts on the future of forensic DNA testing**

John M. Butler · Michael D. Coble · Peter M. Vallone



Forensic Science International: Genetics 3 (2009) 233–241

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

ELSEVIER

Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples

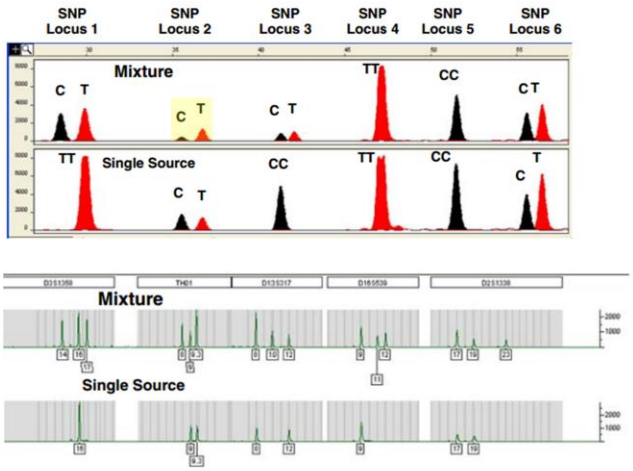
Antoinette A. Westen [a], Anuska S. Matai [a], Jeroen F.J. Laros [b], Hugo C. Meiland [b], Mandy Jasper [a], Wiljo J.F. de Leeuw [a], Peter de Knijff [c], Titia Sijen [a,*]

[a] Department of Biology (R&D), Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
[b] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[c] Forensic Laboratory for DNA Research, Leiden University Medical Center, P.O. Box 9503, 2300 RA Leiden, The Netherlands

## SNP Issues

### Mixture Detection

*Butler 2007*



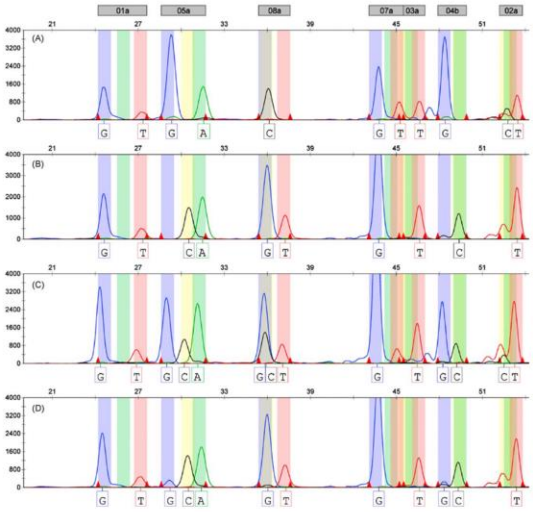

CE based SNP typing data– mixture might be detectable at SNP locus 2, but peak imbalance is common even in single source

CE based STR typing data— mixture is detectable at all loci shown

## SNP Issues

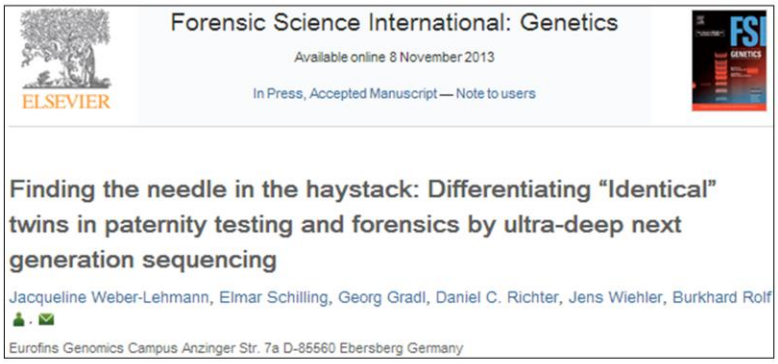
### Mixture Detection

*Westen 2009*




CE based SNP typing data with triallelic SNPs. Two individuals run singly and in 1:1 and 1:8 mixtures. 7 tri-allelic loci were run– at 5 loci we cannot detect a mixture.

## SNP Information



Forensic Science International: Genetics

Available online 8 November 2013

In Press, Accepted Manuscript — Note to users

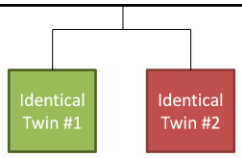
**Finding the needle in the haystack: Differentiating "Identical" twins in paternity testing and forensics by ultra-deep next generation sequencing**

Jacqueline Weber-Lehmann, Elmar Schilling, Georg Gradl, Daniel C. Richter, Jens Wiehler, Burkhard Rolf

Eurofins Genomics Campus Anzinger Str. 7a D-85560 Ebersberg Germany

This article is an interesting application of NGS to distinguish identical twins, which is not possible with current STR technology.

## SNP Information





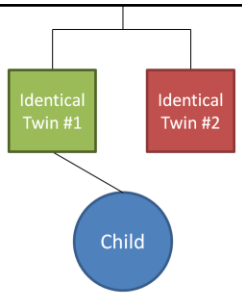
- Criminal paternity example
  - Comparison of twin's sperm DNA genomes

| Adapted from manuscript Table 1 | Chromosome 04 | Chromosome 06 | Chromosome 11 | Chromosome 14 | Chromosome 15 |
|---|---|---|---|---|---|
| | snp188267982 | snp41885722 | snp68781324 | snp103545720 | cnp57884799 |
| Twin #1 sperm | C/T (80/20) | G/A (70/30) | C/T (80/20) | A/G (50/50) | C/T (50/50) |
| Twin #1 buccal mucosa | C/T (80/20) | G/A (75/25) | C/C | A/G (60/40) | C/T (50/50) |
| Twin #1 blood | C/C | A/A | C/C | G/G | C/T (80/20) |
| | | | | | |
| Twin #2 sperm | C/C | A/A | C/C | G/G | C/C |
| Twin #2 buccal mucosa | C/C | A/A | C/C | G/G | C/C |
| Twin #2 blood | C/C | A/A | C/C | G/G | C/C |

Rare mutations occur early after the human blastocyst has split into two, the origin of twins, and that such mutations will be carried on into somatic tissue…

In this mock case, the two twins samples (sperm, buccal & blood) were whole genome sequenced and compared. The hypothesis was that mutations present in one twin's sperm and not the other's might have carried into the germline and thus be present in any offspring. The SNPs shown are present in one twin's sperm and to varying degrees in his buccal and blood, but they are absent from all of the other twin's samples.

## SNP Information




- Criminal paternity example
  - Comparison of twin's sperm DNA genomes

| Adapted from manuscript Table 1 | Chromosome 04 | Chromosome 06 | Chromosome 11 | Chromosome 14 | Chromosome 15 |
|---|---|---|---|---|---|
| | snp188267982 | snp41885722 | snp68781324 | snp103545720 | cnp57884799 |
| Twin #1 sperm | C/T (80/20) | G/A (70/30) | C/T (80/20) | A/G (50/50) | C/T (50/50) |
| Twin #1 buccal mucosa | C/T (80/20) | G/A (75/25) | C/C | A/G (60/40) | C/T (50/50) |
| Twin #1 blood | C/C | A/A | C/C | G/G | C/T (80/20) |
| Child blood | C/T (50/50) | G/A (50/50) | C/T (50/50) | A/G (50/50) | C/T (50/50) |
| Twin #2 sperm | C/C | A/A | C/C | G/G | C/C |
| Twin #2 buccal mucosa | C/C | A/A | C/C | G/G | C/C |
| Twin #2 blood | C/C | A/A | C/C | G/G | C/C |

Rare mutations occur early after the human blastocyst has split into two, the origin of twins, and that such mutations will be carried on into somatic tissue… *and the germline*.

The child shares all these SNPs with one of the twins, thus this twin is expected to be the father.

## SNP Challenges

- High level multiplexing
  - May require pooling
- Mixtures
- New forensic markers
  - Casework use
  - Statistics/reporting

Do we adopt core loci or incorporate SNPs with overlapping roles/purposes to allow for failures?

Do we validate the same way as other markers if we are only using for investigative leads?

Do we validate lab processing/genotyping in a way that allows for failures?

The prediction models are evolving and predictions are often lower than we expect from forensic statistics

# NGS Forensic Applications

SNP Genotyping

- ☑ Sample type amenable to library preparation?
- ☑ Sample type amenable to sequencing platforms?
- ☑ Sample type amenable to bioinformatics?
- ☑ Improvement over current method?

SNPs are completely amenable to NGS typing and bioinformatics. No one method has been adopted by the community for forensic SNP typing– these markers are not yet in common use, so there is no "current method" for comparison.

# Use of NGS for forensic applications

Highly-parallel/high-throughput direct sequencing of forensically relevant targets

- Whole mitochondrial genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Forensically relevant SNPs
  - newer human identity applications
  - biogeographical ancestry, externally visible traits, complex kinship
  - degraded samples, mixtures, low template?
- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'

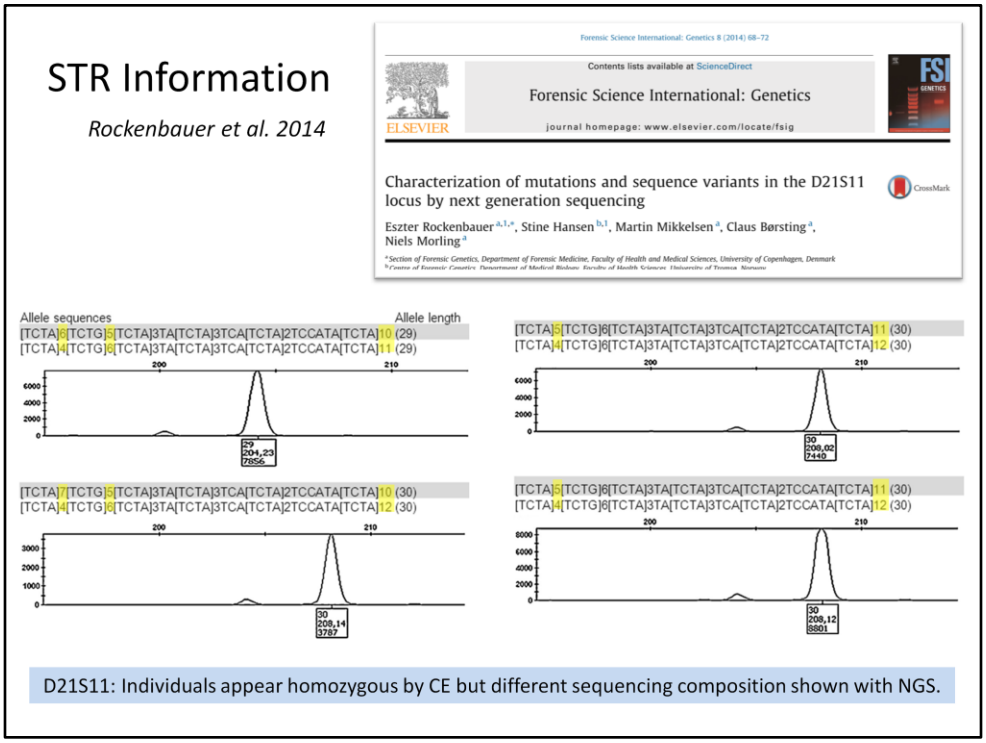
Legacy– new NGS technology needs to be back compatible for databases

What information could be generated by sequencing STRs?

- A. Changes in repeat unit motifs
- B. SNPs/Indels within the repeat regions
- C. Sequences of incomplete repeat units
- D. SNPs in flanking regions

# STR Information

*Rockenbauer et al. 2014*




D21S11: Individuals appear homozygous by CE but different sequencing composition shown with NGS.

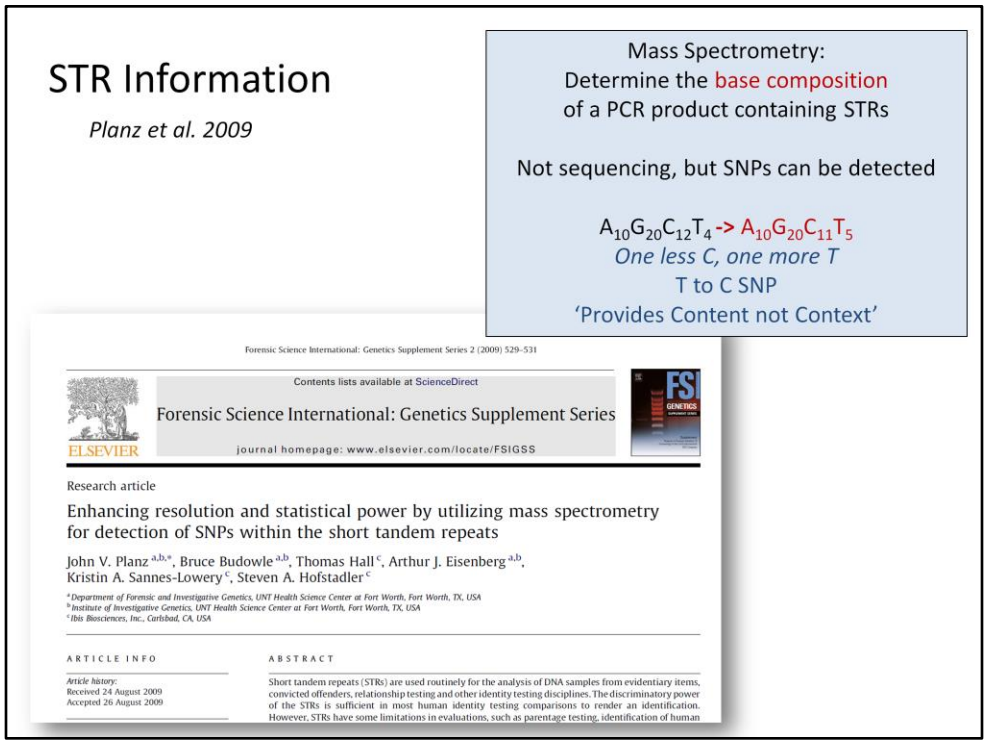
DNA sequences and CE results in four unrelated individuals reported as homozygous by CE, but with different sequence composition as shown with NGS.

## STR Information

## D21S11

| | | | | | |
|---|---|---|---|---|---|
| $[TCTA]_{4-13}$ | $[TCTG]_{3-11}$ | {$[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{6-15}$ | | |
| $[TCTA]_{4-6}$ | $[TCTG]_{5-6}$ | {$[TCTA]_{2-3}$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{8-16}$ | TA | [TCTA] |
| $[TCTA]_{5-11}$ | $[TCTG]_{6-14}$ | {[----] -- $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{9-13}$ | | |
| $[TCTA]_{5-6}$ | $[TCTG]_{5-6}$ | {$[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{5-10}$ | TCA | $[TCTA]_{2-6}$ NNN... |

D21S11 has significant variation of repeat units throughout the complex repeat motif.




Both are 46 bases long, but differ in base content – they will also have unique masses

# STR Information

*Planz et al. 2009*

| Locus | Population | STR only analysis on IBIS T5000 | | | STR-SNP analysis on IBIS T5000 | | |
|---|---|---|---|---|---|---|---|
| | | *n* | Alleles detected | DP | *n* | Alleles detected | DP |
| D13S317 | Caucasian | 182 | 7 | 0.9213 | 181 | 12 | 0.9705 |
| | African Am. | 214 | 7 | 0.8607 | 213 | 12 | 0.9528 |
| | Hispanic | 193 | 7 | 0.9445 | 193 | 13 | 0.9751 |
| D21S11 | Caucasian | 182 | 14 | 0.9540 | 181 | 23 | 0.9780 |
| | African Am. | 214 | 20 | 0.9589 | 213 | 33 | 0.9708 |
| | Hispanic | 193 | 14 | 0.9521 | 193 | 25 | 0.9752 |
| D3S1358 | Caucasian | 182 | 8 | 0.9226 | 181 | 18 | 0.9671 |
| | African Am. | 214 | 8 | 0.8923 | 213 | 18 | 0.9775 |
| | Hispanic | 193 | 8 | 0.8939 | 193 | 18 | 0.9455 |
| D5S818 | Caucasian | 182 | 9 | 0.8432 | 181 | 15 | 0.9260 |
| | African Am. | 214 | 9 | 0.8932 | 213 | 17 | 0.9102 |
| | Hispanic | 193 | 9 | 0.8679 | 193 | 13 | 0.9554 |
| D7S820 | Caucasian | 182 | 8 | 0.9349 | 181 | 15 | 0.9600 |
| | African Am. | 214 | 8 | 0.7 | 213 | 12 | 0.9376 |
| | Hispanic | 193 | 9 | 0.7358 | 193 | 14 | 0.9482 |
| D8S1179 | Caucasian | 182 | 10 | 0.9324 | 181 | 14 | 0.9627 |
| | African Am. | 214 | 10 | 0.9239 | 213 | 19 | 0.9489 |
| | Hispanic | 193 | 9 | 0.9303 | 193 | 16 | 0.9639 |
| vWA | Caucasian | 182 | 10 | 0.9388 | 181 | 22 | 0.9580 |
| | African Am. | 214 | 11 | 0.9403 | 213 | 26 | 0.9766 |
| | Hispanic | 193 | 7 | 0.9108 | 193 | 16 | 0.9305 |

Average number of alleles nearly doubles
Discriminating power increases 3.5–5% per locus

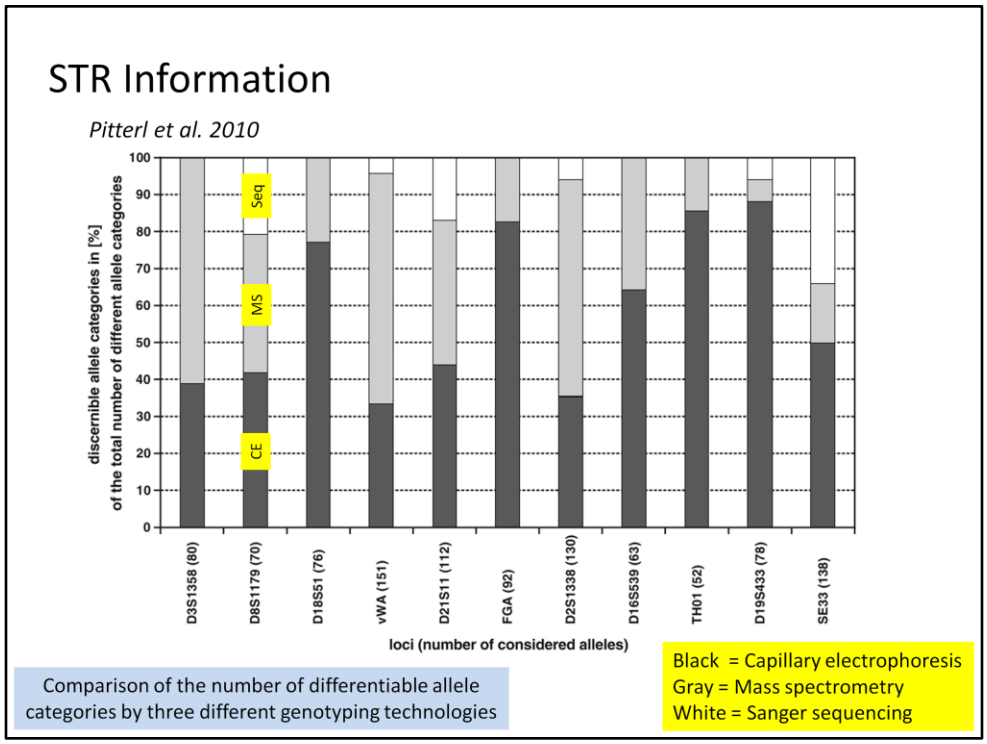
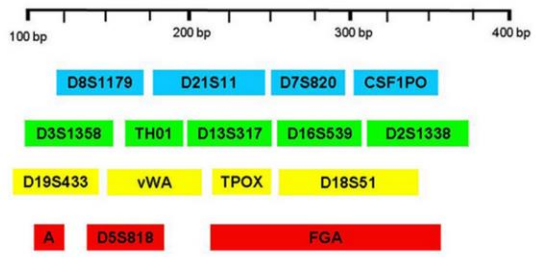



Fig. 1 Comparison of the number of differentiable allele categories by three different genotyping technologies. For each locus, a representative number of alleles (numbers in
parentheses) were analyzed by electrophoresis, ICEMS, and Sanger sequencing.

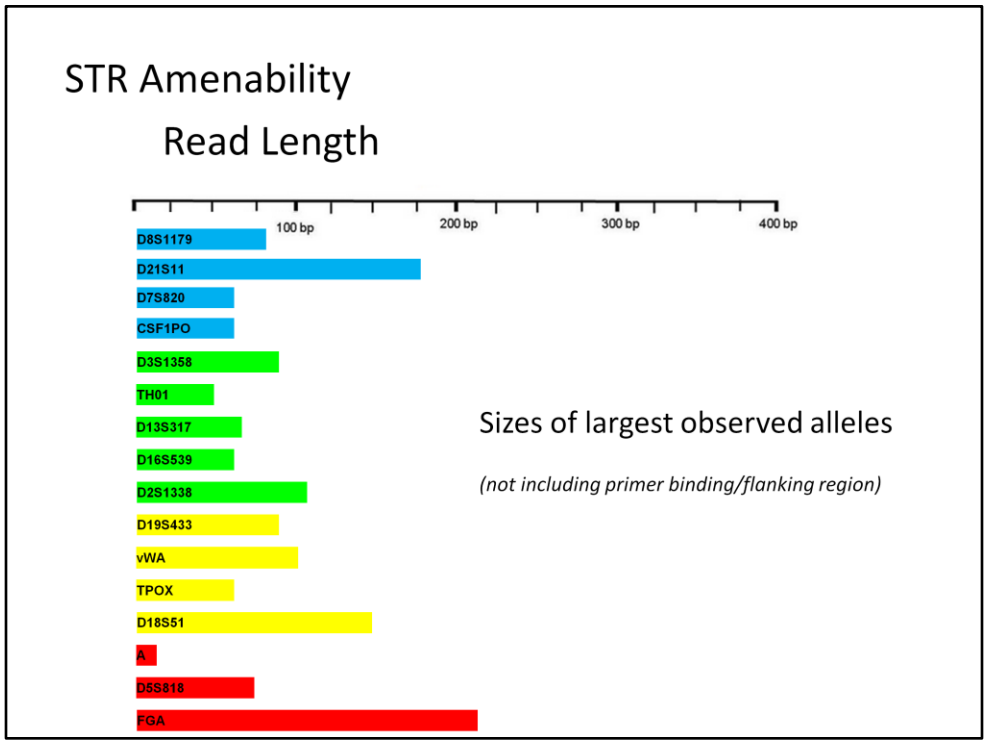
The number of allele categories discernible via electrophoresis (black) and ICEMS (gray) is compared to
the number of differentiable allele classes by sequence analysis and expressed in percent. Sequence analysis (white) corresponds to 100% of discernible allele categories

## STR Amenability

### Read Length



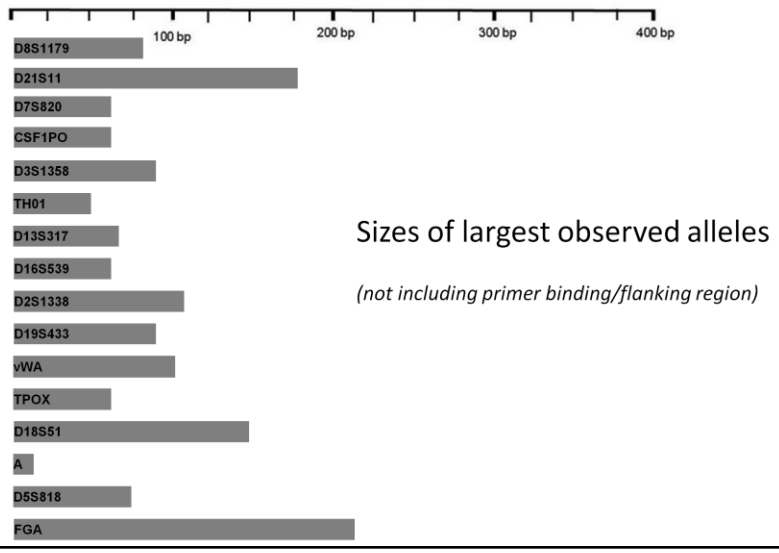

Example from identifiler




The STR loci no longer need to be separated based on size, as was the case with CE genotyping.

## STR Amenability

### Read Length




Sizes of largest observed alleles

*(not including primer binding/flanking region)*

They also no longer need to be fluorescently labeled.

What is the most important limiting factor in reducing the size of these amplicons?

A. Secondary structures are more common in shorter sequences.

B. Primer dimers are more likely with shorter sequences.

C. Primer binding regions are subject to constraints such as GC content or SNPs.

D. Sequencing chemistry performs better with longer amplicons.

Why might 150bp paired reads be problematic for STRs?

A. An STR amplicon can only be read in one direction.

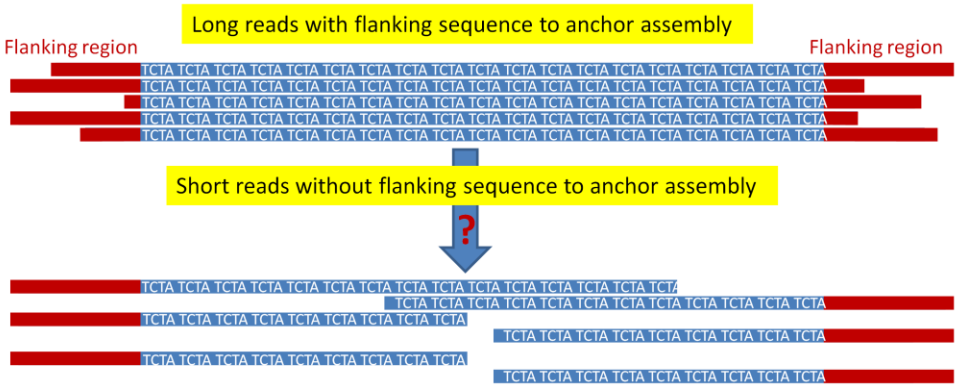
B. Each read must extend beyond the repeat region.

C. The required indexing primers cannot bind to repeat regions.

D. There is no reference sequence for STR loci.

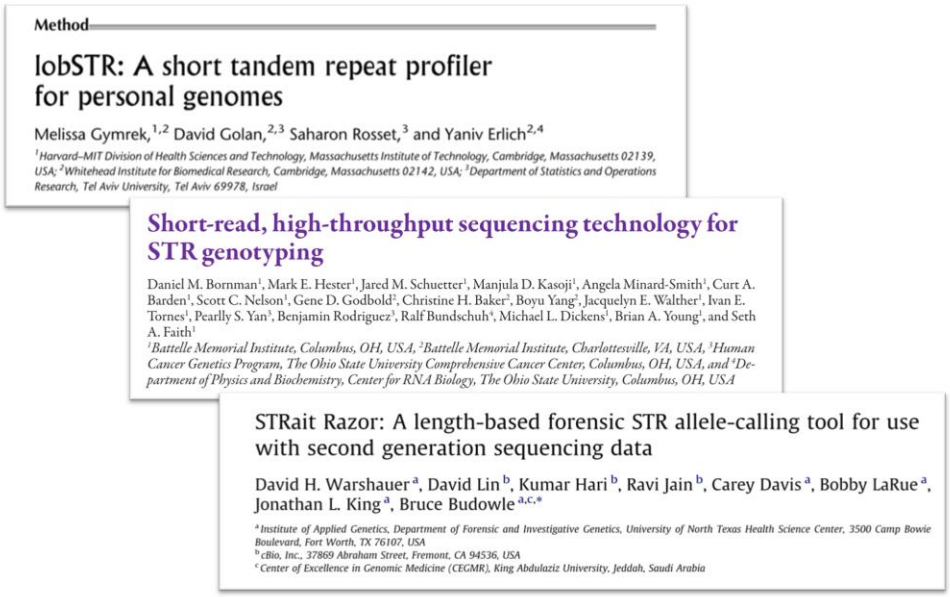
# STR Amenability

## Read Length




**Both MiSeq (2x300) and PGM (400bp) can read completely through STR repeat regions**

# STR Amenability - Bioinformatics



Method

## lobSTR: A short tandem repeat profiler for personal genomes

Melissa Gymrek,[1,2] David Golan,[2,3] Saharon Rosset,[3] and Yaniv Erlich[2,4]

[1]Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [2]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [3]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel

## Short-read, high-throughput sequencing technology for STR genotyping

Daniel M. Bornman[1], Mark E. Hester[1], Jared M. Schuetter[1], Manjula D. Kasoji[1], Angela Minard-Smith[1], Curt A. Barden[1], Scott C. Nelson[1], Gene D. Godbold[2], Christine H. Baker[2], Boyu Yang[3], Jacquelyn E. Walther[1], Ivan E. Tornes[1], Pearlly S. Yan[3], Benjamin Rodriguez[3], Ralf Bundschuh[4], Michael L. Dickens[1], Brian A. Young[1], and Seth A. Faith[1]

[1]Battelle Memorial Institute, Columbus, OH, USA, [2]Battelle Memorial Institute, Charlottesville, VA, USA, [3]Human Cancer Genetics Program, The Ohio State University Comprehensive Cancer Center, Columbus, OH, USA, and [4]Department of Physics and Biochemistry, Center for RNA Biology, The Ohio State University, Columbus, OH, USA

## 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] eBio, Inc., 37869 Abraham Street, Fremont, CA 94536, USA
[c] Center of Excellence in Genomic Medicine (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

# STR Amenability - Bioinformatics

Gymrek et al. 2012

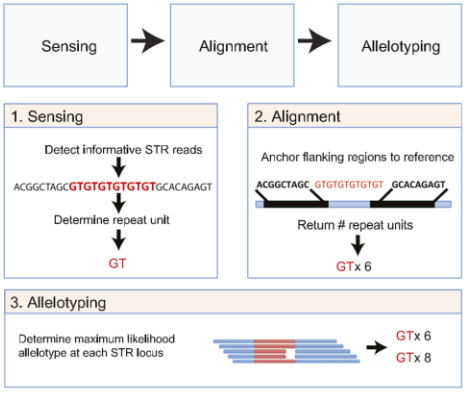
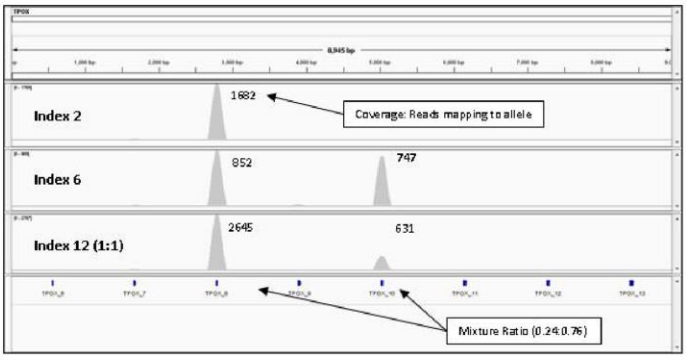



**Figure 1.** lobSTR algorithm overview. lobSTR consists of three steps. The sensing step detects informative STR reads and determines their repeat motif. The alignment step maps the STRs' flanking regions to the reference. The allelotyping step determines the STR alleles present at each locus.

This method first detects the repeat motif and then anchors to a reference.

## STR Amenability - Bioinformatics

Bornman et al. 2012



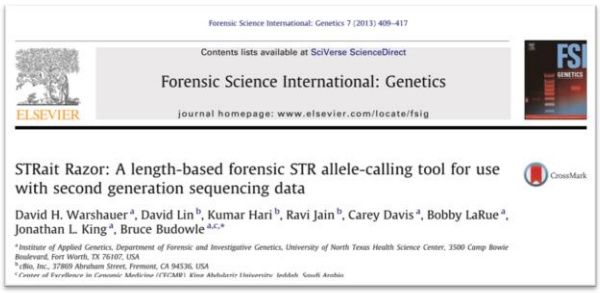

By aligning the data to an silico reference sequence, the genotype can be determined, and known sequence variants can also be detected. New repeat motifs (not included in the reference sequence) may not be detected.

# STR Amenability - Bioinformatics



AGCTCGATCGCGCTAGTTTAGAGCattgactctatcGATAGATAGATAGATAGATAGATAGATAGATAcctgatagtattGATTCTATCGATATATCCTAGCTA
~~ATAGATAGATAGATAGATAcctgatagtattGATTCTATCGATATATCCTAGCTAGAATCTCGGTGATCTAGCTATTTATCCGCTAGCCATATGTGCGATCGAATC~~
~~GCCTAGCTAGCTTAGCTATATCTAGATAGCTAGCTAGCTTTCGCGAAAGCTCGATCGCGCTAGTTTAGAGCattgactctatcGATAGATAGATAGATAGATAGA~~
ATCGCGCTAGTTTAGAGCattaactctatcGATAGATAGATAGATAGATAGATAGATAGATAcctgacagtattGATTCTATCGATATATCCTAGCTAGAATCT

1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.



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

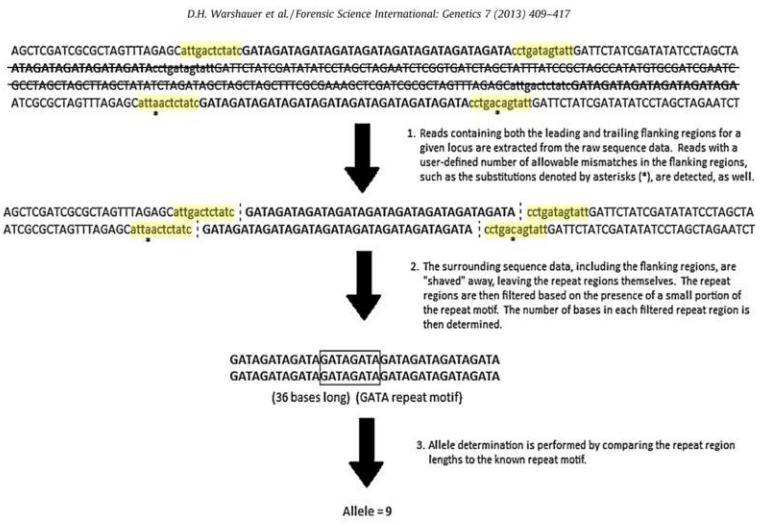
STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

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

This method looks for flanking regions (lower case highlighted), and only uses sequences where both 5’ and 3’ end flanking regions are present.

# STR Amenability - Bioinformatics



## STR Amenability - Bioinformatics

### Bioinformatics are key to identifying sub-alleles

| | | | | | |
|---|---|---|---|---|---|
| $[TCTA]_{4-13}$ | $[TCTG]_{3-11}$ | {$[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{6-15}$ | | |
| $[TCTA]_{4-6}$ | $[TCTG]_{5-6}$ | {$[TCTA]_{2-3}$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{8-16}$ | TA | [TCTA] |
| $[TCTA]_{5-11}$ | $[TCTG]_{6-14}$ | {[----] -- $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{9-13}$ | | |
| $[TCTA]_{5-6}$ | $[TCTG]_{5-6}$ | {$[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA} | $[TCTA]_{5-10}$ | TCA | $[TCTA]_{2-6}$ NNN... |

lobSTR and STRaight Razor methods give length based genotype.

In silico reference can give known sub alleles.

Ideal method wouldn't be constrained to a ladder and would return sequence variants & changes in repeat motif.

## STR Challenges

- Bioinformatics
  - Single source
  - Mixtures
- Nomenclature
- Statistics
  - Subvariant population databases
- Reporting
- Searching

The bioinformatics are challenging even for single source samples.

Nomenclature– develop a new system for denoting sub alleles? Use entire string for database searching? These are questions the community would need to address prior to routinely generating sequencing data from STRs.

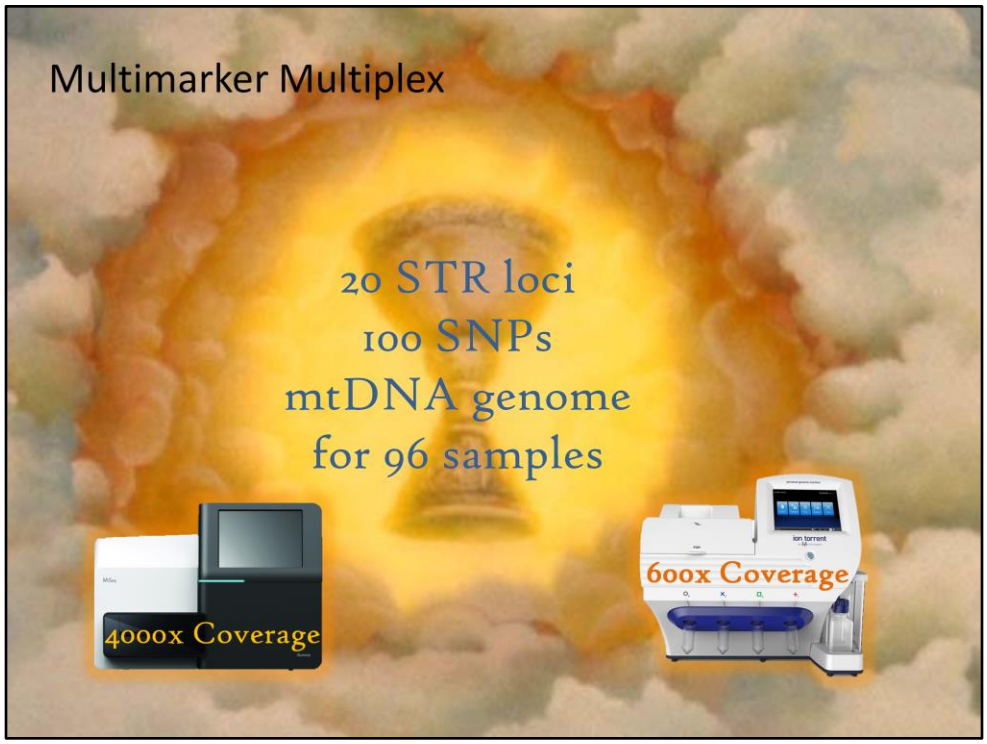
In order to use the sequence variant data in forensic statistics, we need population databases with sequencing data.

# NGS Forensic Applications

STR Sequencing

- ☑ Sample type amenable to library preparation?
- ☑ Sample type amenable to sequencing platforms?
- ☑ Sample type amenable to bioinformatics?
- ☒ Improvement over current method?

The library prep and read length are amenable to STRs at this point. The bioinformatics need improvement. As existing CE-based STR processing is so streamlined, NGS would not currently be an improvement.




This is what a multimarker multiplex could look like, and the rough estimate of theoretical coverage with the current technology (2x300 v3 on the MiSeq and 318 chip on the PGM).