Genomic Security
(Lest We Forget)

Gene Tsudik
CS@UCI

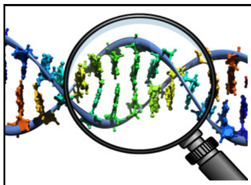
DISCLAIMER

I am:

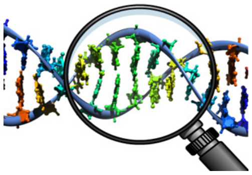
- A researcher in: security, privacy, applied crypto

I am **not**:

- An expert in: genomics, genetics, bioinformatics, statistics, ML, and much of everything else



Basics



- **Genome**
  - A complete blueprint of an organism
  - At least one copy in almost all cells
  - Encoded in DNA: double stranded polymer of nucleotides:

    A, C, G, T
  - In humans, 3.2 Billion nucleotides (in 23 chromosome pairs)
- **Whole Genome Sequencing (WGS)**
  - Process of determines complete DNA sequence of an organism's genome

NOTE: the rest of this talk is blatantly ***specieist***

3

Storage/Representation

- Full hypothetical: about 720 Mbytes
- Raw sequencer output: >200 Gbytes
  - Short reads: many redundant "short reads"
  - FASTQ file format (ASCII)
- Variances/differences: about 130 Mbytes
  - Based on a fixed reference genome: **GRCh38.p10**
  - Uses above short reads to align
  - Captures roughly 0.1% difference ($3.2*10^6$)
  - VCF file format (ASCII)
  - One SNP (single-nucleotide polymorphism) per data line

VCF: one SNP example

```
#CHROM  POS    ID          REF  ALT  QUAL  FILTER  INFO                      FORMAT       NA00001         NA00002         NA00003
20      14370  rs6054257   G    A    29    PASS    NS=3;DP=14;AF=0.5;DB;H2   GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
```

http://samtools.github.io/hts-specs/VCFv4.2.pdf

WGS Progress

- **Chronology:**
  - – 1970s: DNA sequencing starts
  - – 1990: The Human Genome Project starts
  - – 2003: First human genome sequenced
  - – 2010: Race for 1,000 genomes ends
- **Cost/genome:**
  - – $3B: The Human Genome Project
  - – $250K: Illumina (2008)
  - – $5K: Complete Genomics (2009), Illumina (2011)
  - – $1K: Life Technologies (2012), Oxford Nanopore (2013)

  Now – race towards $100

Now What?

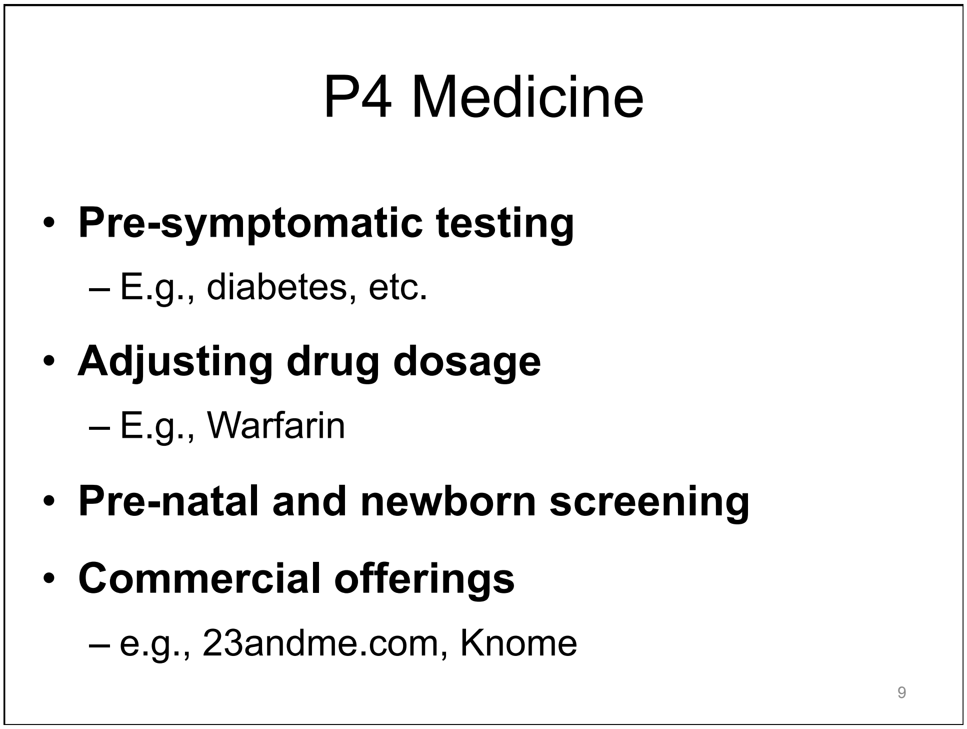
- **Ubiquitous affordable WGS:** a promise for the very near future
- **The Good News**
  - More efficient/powerful/cost-effective genomic tests
    - Improving and reducing costs of healthcare
  - Facilitating “P4 Medicine”: **P**redictive, **P**reventive, **P**articipatory, and **P**ersonalized
  - Enabling Genome-Wide Association Studies (GWAS)
- **The Bad News**
  - Numerous privacy, security and ethical concerns

7

P4 Medicine

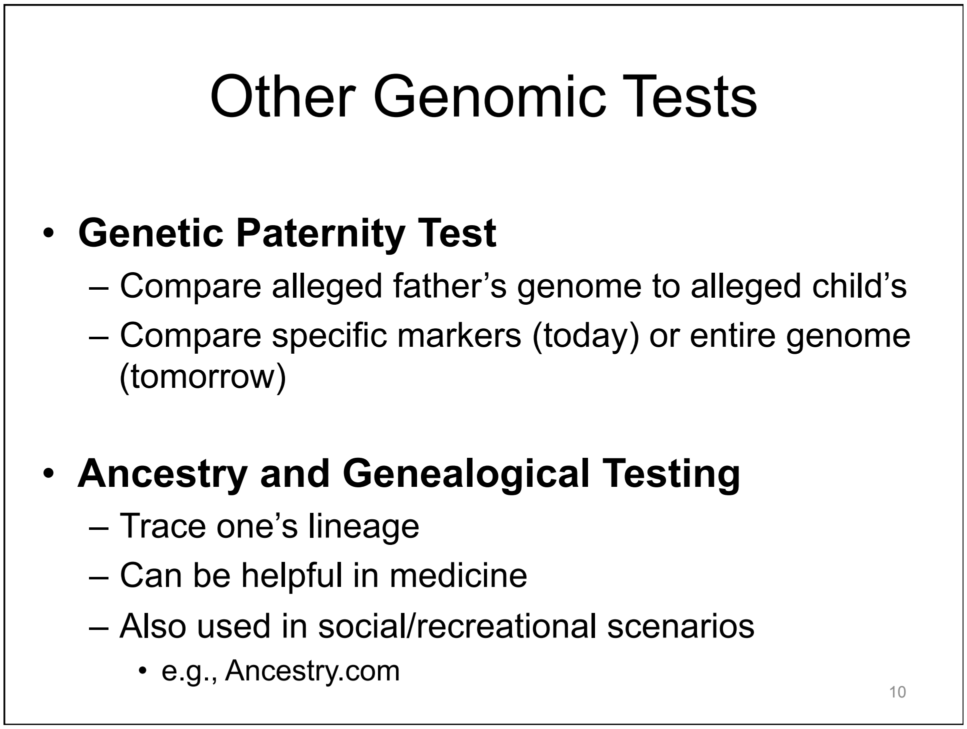
- **Diagnosis & treatment tailored to a specific patient’s genome**
  - Better understanding of the disease
  - More effective medication
- **A few examples:**
  - **tmpt** mutations tested before treating child leukemia
  - **brca1/brca2** correlated to breast and ovarian cancers
  - **hla-B*** tested for HIV drug
  - **erbB2** tested in relation to breast, lung, colorectal cancer

8



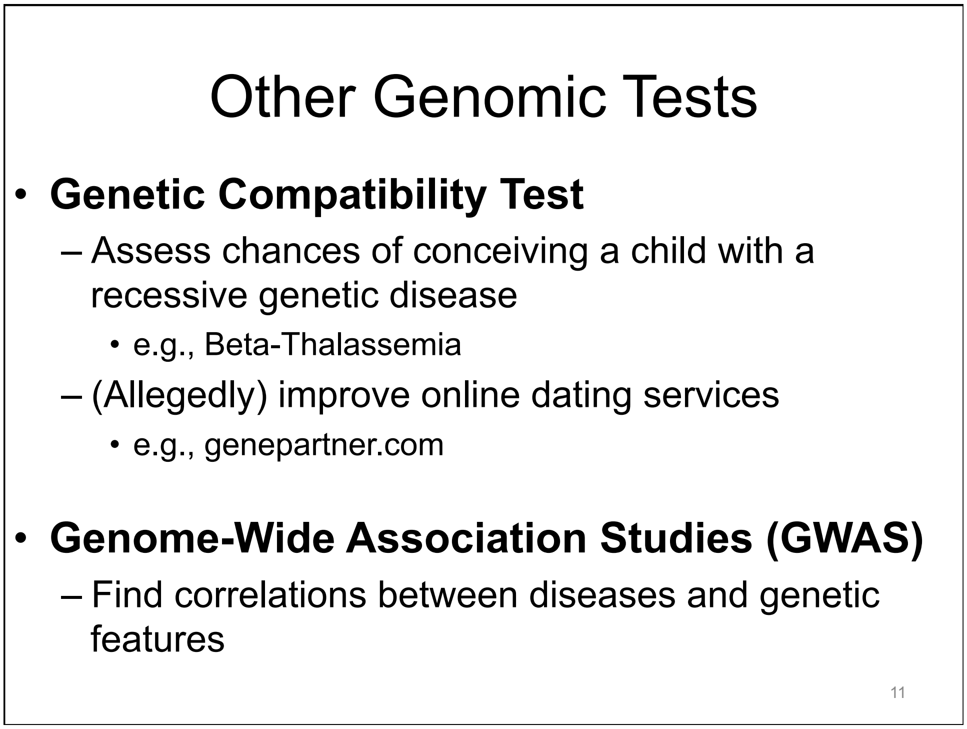
P4 Medicine

- **Pre-symptomatic testing**
  - E.g., diabetes, etc.
- **Adjusting drug dosage**
  - E.g., Warfarin
- **Pre-natal and newborn screening**
- **Commercial offerings**
  - e.g., 23andme.com, Knome



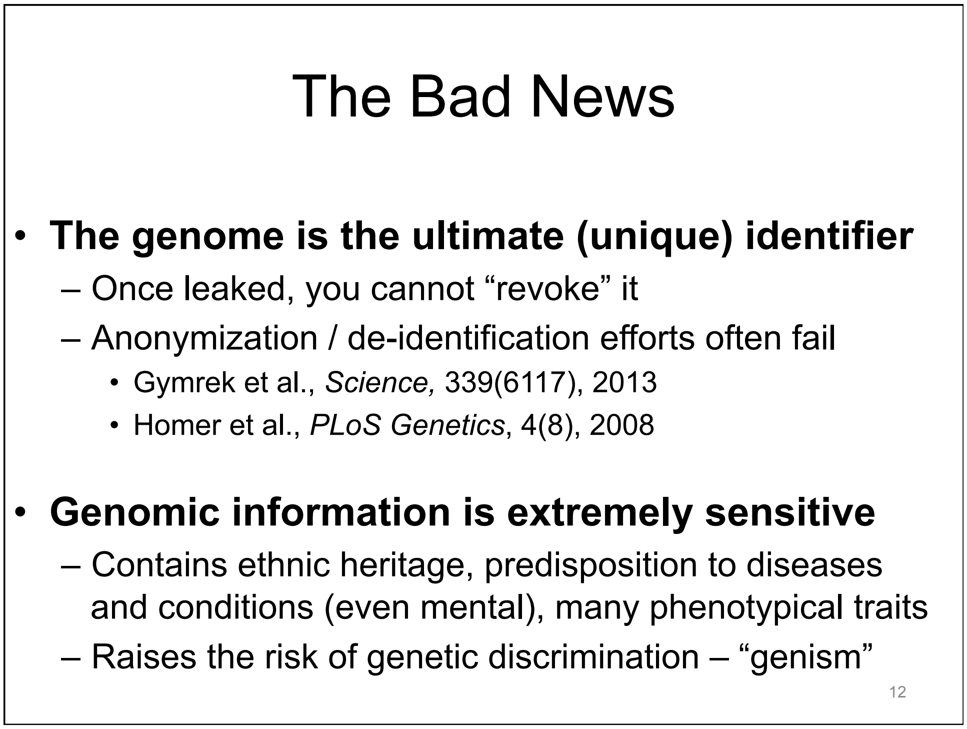
Other Genomic Tests

- **Genetic Paternity Test**
  - Compare alleged father’s genome to alleged child’s
  - Compare specific markers (today) or entire genome (tomorrow)
- **Ancestry and Genealogical Testing**
  - Trace one’s lineage
  - Can be helpful in medicine
  - Also used in social/recreational scenarios
    - e.g., Ancestry.com



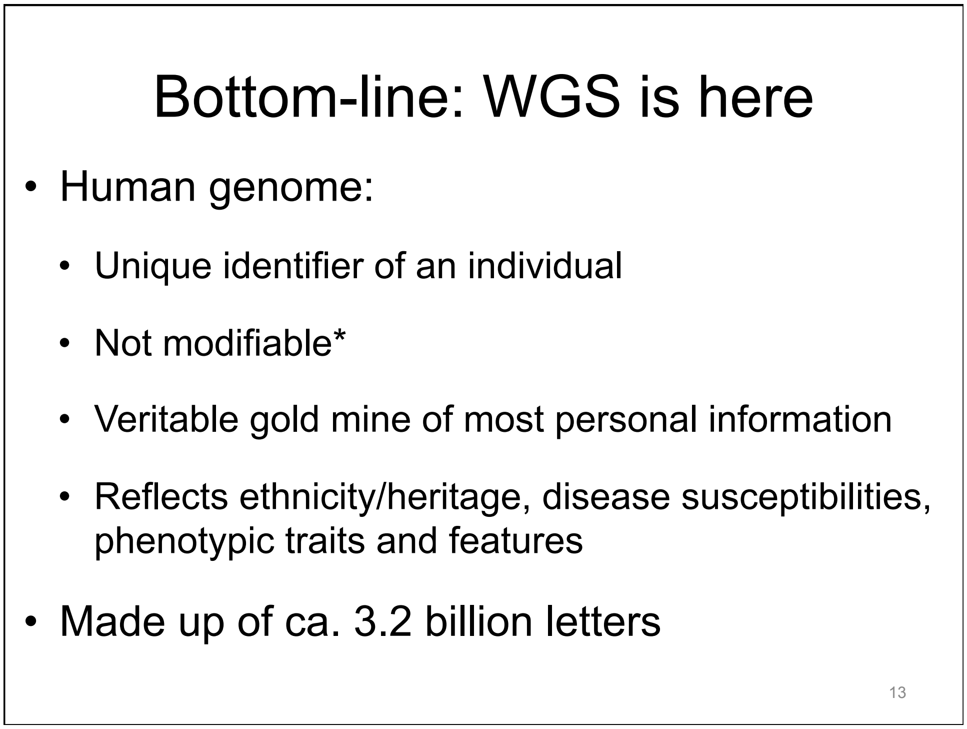
Other Genomic Tests

- **Genetic Compatibility Test**
  - Assess chances of conceiving a child with a recessive genetic disease
    - e.g., Beta-Thalassemia
  - (Allegedly) improve online dating services
    - e.g., genepartner.com
- **Genome-Wide Association Studies (GWAS)**
  - Find correlations between diseases and genetic features



The Bad News

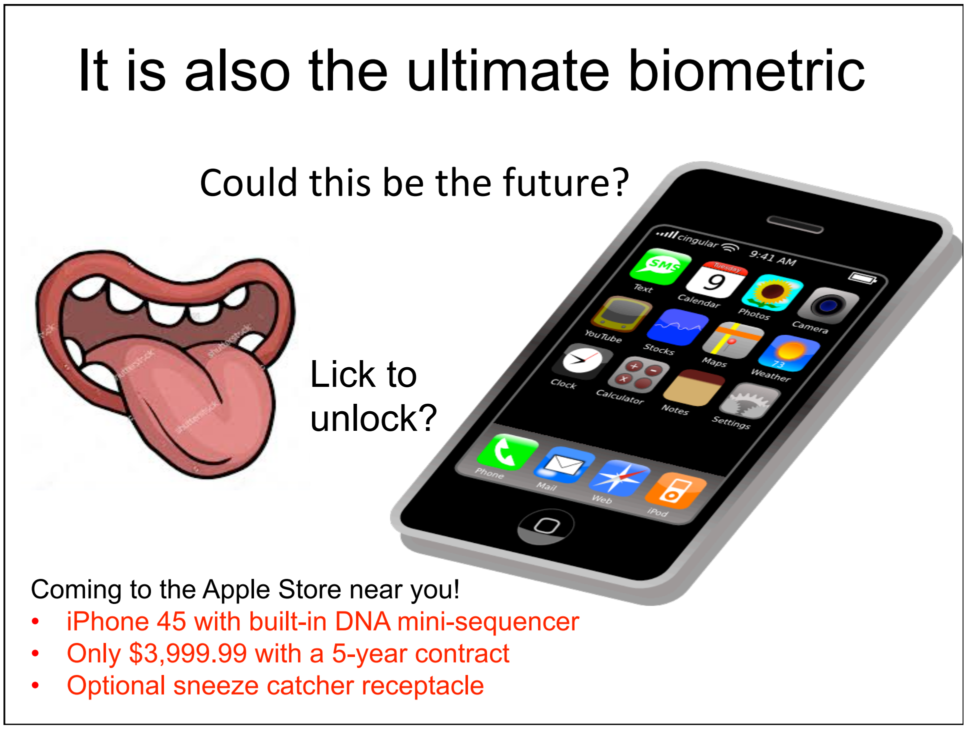
- **The genome is the ultimate (unique) identifier**
  - Once leaked, you cannot “revoke” it
  - Anonymization / de-identification efforts often fail
    - Gymrek et al., *Science,* 339(6117), 2013
    - Homer et al., *PLoS Genetics*, 4(8), 2008
- **Genomic information is extremely sensitive**
  - Contains ethnic heritage, predisposition to diseases and conditions (even mental), many phenotypical traits
  - Raises the risk of genetic discrimination – “genism”



Bottom-line: WGS is here

- Human genome:
  - Unique identifier of an individual
  - Not modifiable*
  - Veritable gold mine of most personal information
  - Reflects ethnicity/heritage, disease susceptibilities, phenotypic traits and features
- Made up of ca. 3.2 billion letters

13



It is also the ultimate biometric

Could this be the future?

Lick to unlock?

Coming to the Apple Store near you!

- iPhone 45 with built-in DNA mini-sequencer
- Only $3,999.99 with a 5-year contract
- Optional sneeze catcher receptacle

It is also a curse

That keeps on cursing…

Once revealed, can’t be changed or revoked

Includes information about:
- Oneself
- Ancestors
- Siblings
- Progeny

No other biometric is like that!

Privacy dominates the spotlight!

- Threats appear to be almost immediate, spectacular and terrifying
- Leakage can be direct or indirect, e.g., surname or location inferencing
- Leakage can be massive, e.g., hacked genomic data-banks
- Attack classes:
  - **Large-Scale (impersonal)**: by cyber-criminals, pharmaceuticals, insurance companies, nations
  - **Targeted (personal)**: by competitors, litigants, “friends”, relatives, nations
- Progress has been made against large-scale attacks
- But, new ones keep popping up
- Targeted attacks seem very hard (perhaps impossible) to mitigate

WHY?

We constantly shed DNA material

- Hair (with root)
- Saliva
- Blood
- Skin cells
- Nail clippings (possibly)
- …
- and so on, and so forth

There ain’t no cure for the focused attack

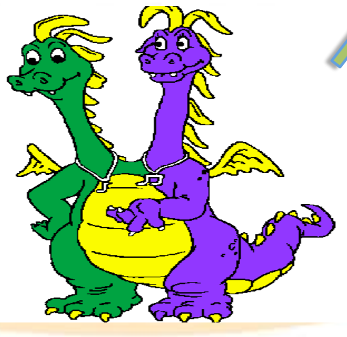
Not even a full-body condom…
And, let’s not forget exhibitionist idiots

https://genomeprivacy.org/

WHAT ABOUT GENOMIC SECURITY?

WHY IT HASN’T RECEIVED MUCH ATTENTION?

Security



Privacy

Hypothetical Scenario (1)

- Alice gets her genome sequenced by a licensed Sequencing Laboratory (SL)
- Alice's fully sequenced digitized genome is stored on her personal device
- Alice's genome is then modified by:
    - Malware
    - Directly (physically) by adversary
    - Alice herself
- Now what?

Hypothetical Scenario (2)

- Alice goes to the doctor who treats her condition (e.g., cancer) using personalized medicine. Wrong medicine is administered.
- Alice is admitted to a hospital. Wrong treatment is administered.
- Alice takes part in a parentage test. Wrong outcome!
- Alice submits genomic information to dating app. Gets paired up fraudulently. The horror! ☺

Security Issues

- Who sequenced the genome?
  - Can that entity be trusted?
  - Who/how certifies this entity?
- Was sequencing done “by the book”?
  - Has the owner consented? or
  - Was the sample otherwise legally obtained?
  - Evidence? Raw data preservation?

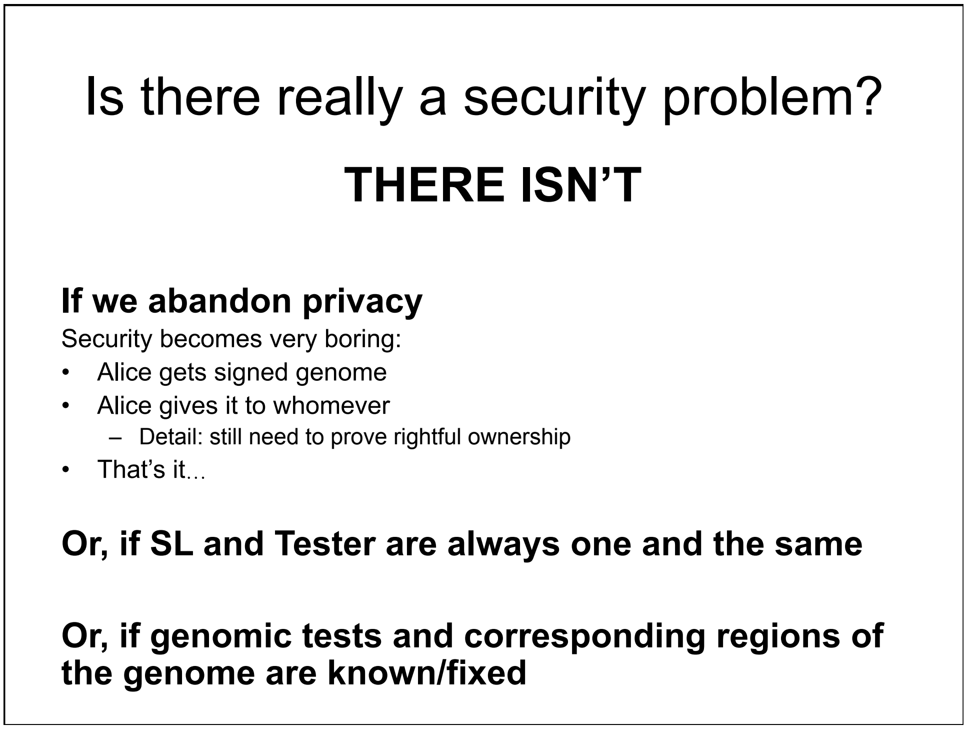
- Has the genome been modified?
- Does the genome belong to its claimed owner?
  - How to authenticate the owner?

- Who has the rights/reasons to “see” which portions of the genome?
  - How to authorize, certify, authenticate, etc., such entities?

Setting, Assumptions, etc.

SL	Licensed sequencing laboratory
Alice	A human being
Tester	Entity given some or all of Alice’s genome • Medical: hospital, clinic, doctor • Legal: court-appointed lab • Social: ancestry or dating app

CL	Cloud service provide
AUTH	“Higher authority”, e.g., FDA



Is there really a security problem?

THERE ISN'T

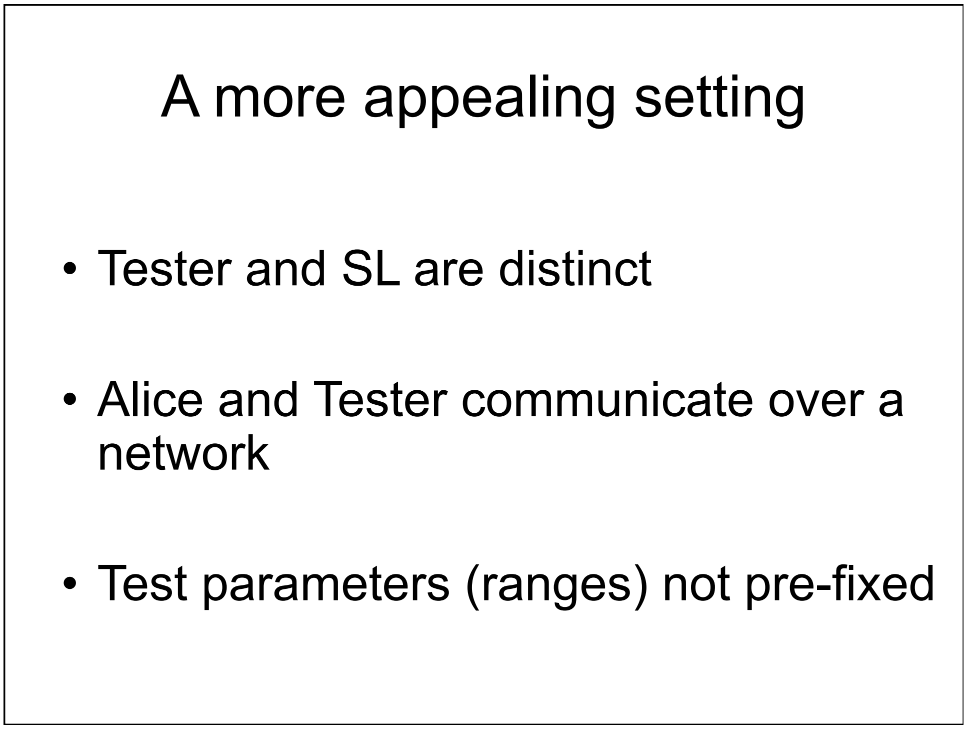
If we abandon privacy

Security becomes very boring:

- Alice gets signed genome
- Alice gives it to whomever
  - Detail: still need to prove rightful ownership
- That's it…

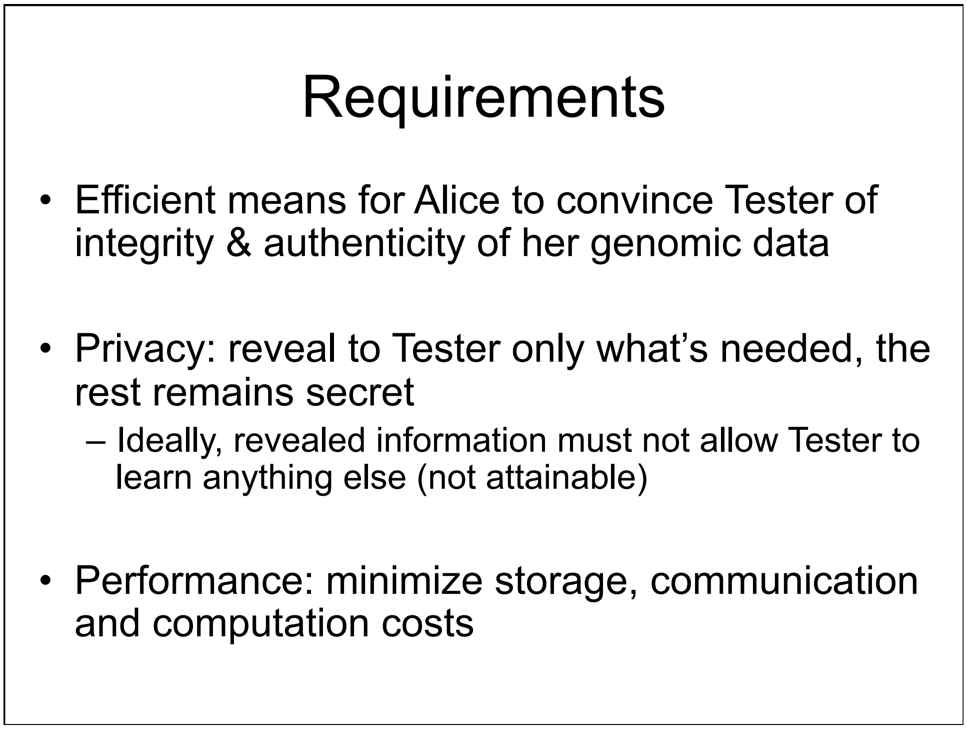
Or, if SL and Tester are always one and the same

Or, if genomic tests and corresponding regions of the genome are known/fixed



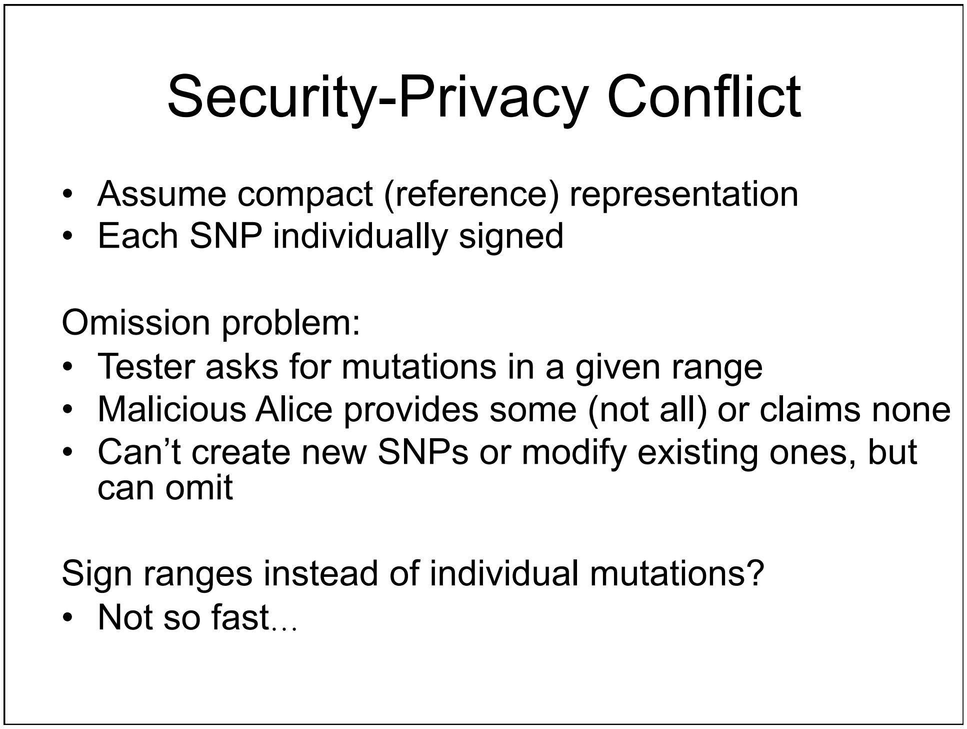
A more appealing setting

- Tester and SL are distinct
- Alice and Tester communicate over a network
- Test parameters (ranges) not pre-fixed



Requirements

- Efficient means for Alice to convince Tester of integrity & authenticity of her genomic data
- Privacy: reveal to Tester only what's needed, the rest remains secret
  - Ideally, revealed information must not allow Tester to learn anything else (not attainable)
- Performance: minimize storage, communication and computation costs



Security-Privacy Conflict

- Assume compact (reference) representation
- Each SNP individually signed

Omission problem:

- Tester asks for mutations in a given range
- Malicious Alice provides some (not all) or claims none
- Can't create new SNPs or modify existing ones, but can omit

Sign ranges instead of individual mutations?

- Not so fast…

EXAMPLE

POS	…	…	…	Y’	Y*	Y”	…	…	…
SNP	…	…	…	C	A	T	…	…	…
sig				σ'	σ*	σ”			

- Tester asks for segment of size X, starting at position Y

  Y>Y’,   Y<Y*,   Y+X<Y”
- Alice has only one SNP in that range: A at Y*
  - Can provide **[Y*,A, σ*]**, or not…(claim no mutations)
  - How to prove absence of other SNPs in requested range?

Similar to completeness in database range query reply

EXAMPLE (contd.)

POS	…	…	…	Y’	Y*	Y”	…	…	…
SNP	…	…	…	C	A	T	…	…	…
SIG				σ'	σ*	σ”			

- Signatures are linked
- No more cheating
- But, Alice would reveal Y’ and Y” along with Y* (plus sig-s)
- Distances: Y-Y’, and Y”-(Y+X) can be VERY LARGE
- Possibly lots of extra information leaked

- The same would hold for other ADS representations, e.g., MHT

How to avoid leakage?

- Revert to full representation…
- Storage is getting cheaper and cheaper
- Alice can store her own genome

And then?

- Sign DNA segments (of what size?)
- Sign each base-letter individually (most flexible)

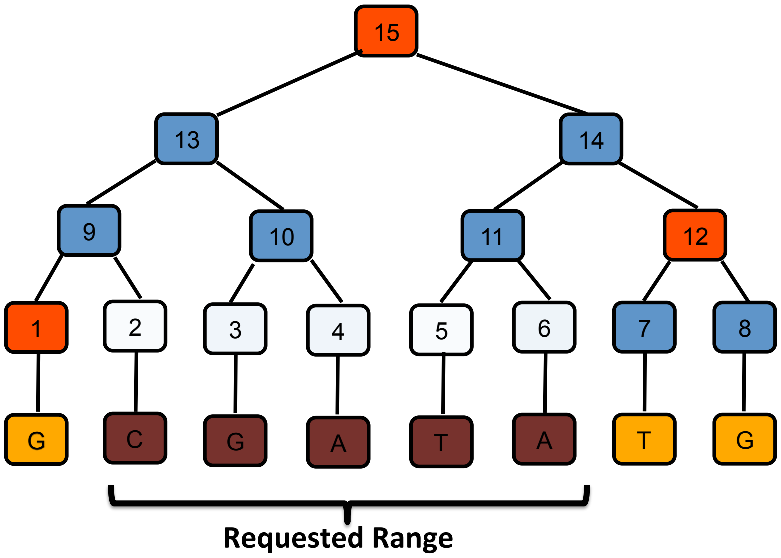
Overhead…

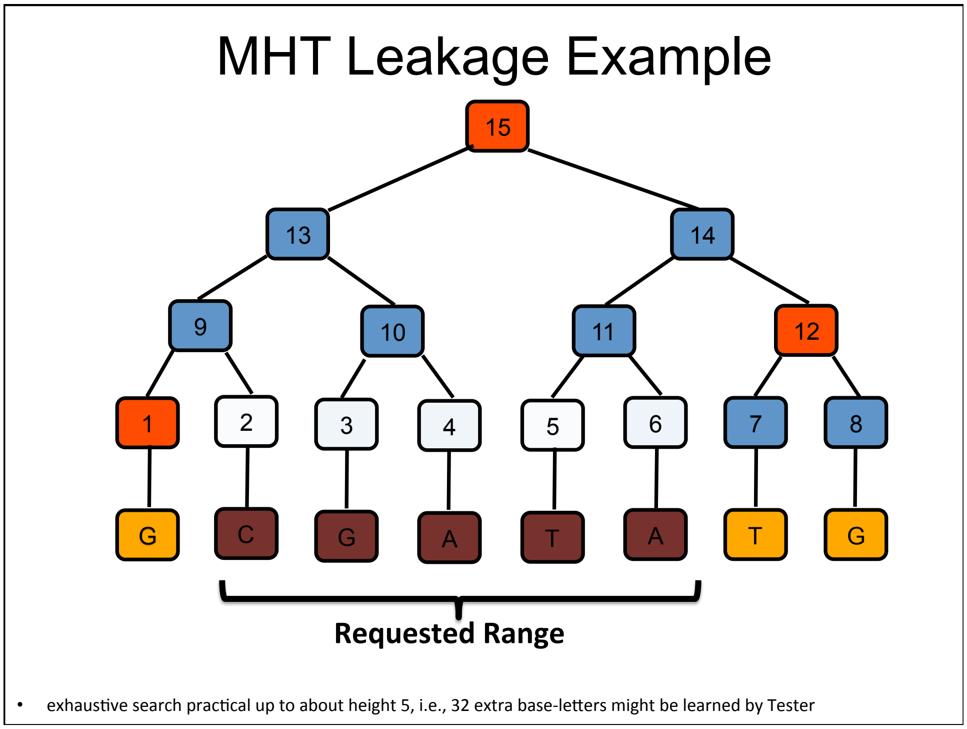
- Signing → not a problem (SL can do it!)
- Extra bits per base-letter: 224 ECC, 2048 RSA
- Transmission and/or verification optimizations:
  - Batch signatures, e.g., w/FDH-RSA, BGR (EC'98)
  - Condensed signatures, e.g., MNT (NDSS'04)
  - Aggregated signatures, e.g., BGLS (EC'03)

Merkle Hash Tree

- Phillips screwdriver equivalent ☺
- SL builds tree with base-letters as leaves
- Signs root
- Height ca. 30
- Storage/computation trade-off for Alice
- Low comp. costs for Tester
  - bunch of hashes + 1 sig ver-n
- Could also use other ADS-s, e.g., skip-lists

Merkle Hash Tree (contd)

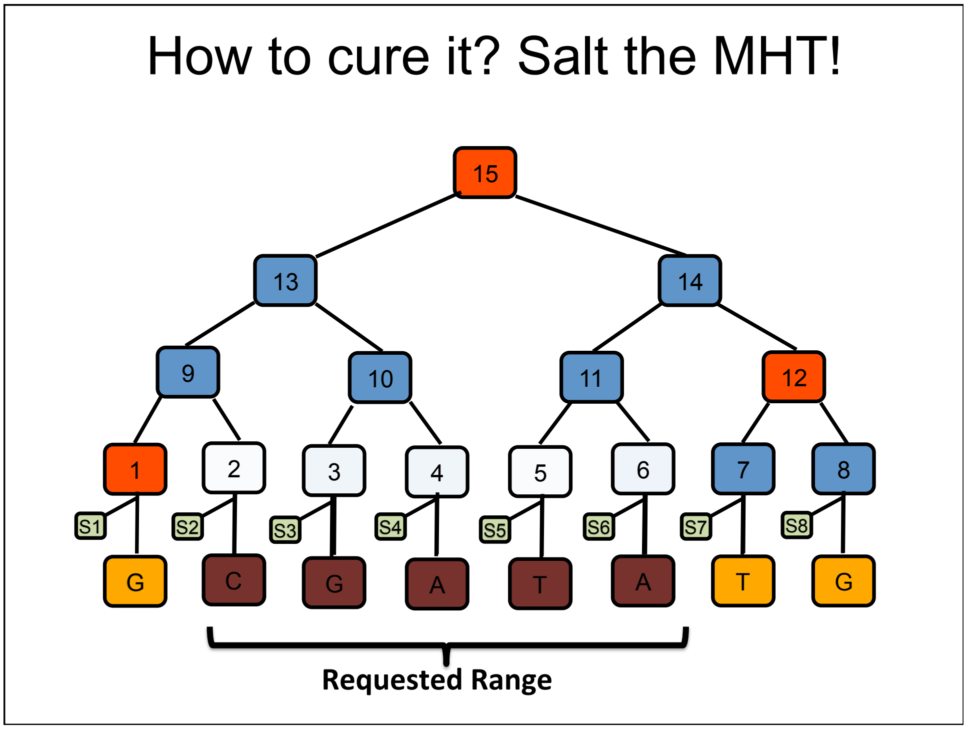




MHT Leakage Example

Requested Range

- exhaustive search practical up to about height 5, i.e., 32 extra base-letters might be learned by Tester



How to cure it? Salt the MHT!

Requested Range

Salted MHT

- Salted by SL at creation time
- Salts generated from master key via PRF
- Key given to Alice
- Salts for requested leaves revealed to Tester

More generally:

- Redactable signatures concept
  - CT-RSA'02, ICISC'01

DSAC

- Signature Aggregation & Chaining
- Given sequence: $L_1,\ldots,L_N$, SL computes, for $0<i<N$:
  $R0 = s_0$

$$Ri = [\ L_i,\ i,\ s_i,\ H(R_{i-1},\ s_{i-1})\ ],\quad \sigma_i = Fsig\ (Ri\ )$$

where:

- Fsig() – hash-and-sign signature function
- $s_i,\ldots,s_{iN}$ – pseudo-random salts (needed as in MHT)
- H() – hash function

DSAC (contd.)

Tester asks for base-letters in range [i,j]

Alice provides:

1. $\{L_i,\ldots,L_j\}$ and $\{s_i,\ldots,s_j\}$
2. $H(R_{i-1}, s_{i-1})$
3. σ_j

- Very low verification cost!
- Low comm. cost

Are we done?

Not yet… only if we're happy with the full representation

Ideally:

SL signs **reference** representation, such that Alice can:

- redact arbitrary portions, and
- efficiently prove that ranges requested by Tester are fully represented by combination of: (1) reference genome and (2) non-redacted portions, signed by SL

Need progress on redactable signatures and techniques akin to group signature revocation

ALSO: What if Alice wishes to remain anonymous wrt Tester?

So…

- Is genomic security underappreciated?
- Is it important?
- Is it research-worthy?

For further info, see:

∅

This is the slide where the invited talk speaker usually lists self-citations, tastefully ornamented with other references, so as not to appear blatantly self-important.

Shukran!

شكرا...