Genetic Variation and Whole Genome Sequencing

Christopher E. Mason
Assistant Professor
Department of Physiology and Biophysics,
The Feil Family Brain and Mind Research Institute (BMRI) &
The Institute for Computational Biomedicine at
Weill Cornell Medical College and the
Tri-Institutional Program on Computational Biology and Medicine
Fellow of the Information Society Project, Yale Law School
May 14th, 2014
@mason_lab
Genome Sequencing

a) Sequencing affected individuals in families
Sequencing affected relatives
Identify shared variants that are likely to cause disease

Genotype candidate variants
Look for co-segregation with disease in family

b) Extreme-trait sequencing
Whole-genome sequencing
Rare variant enriched in population with extreme phenotype

Targeted genotype
Rare variant associates with phenotype in larger population

Cirulli ET& Goldstein DB. Nature Review Genetics, 2010
A Revolution of Personalized (Public) Medicine
Sharing = Caring
<Pre-Womb>
Womb
Life
Tomb
Post-Tomb
Happy to give no deleterious BRCA1 alleles to the next generation
Sharing
=
Pre-Caring
Carrier Screening Becoming more Common

[Diagram showing inheritance of Cystic Fibrosis (CF).]

- Father (Carrier of CF Gene)
- Mother (Carrier of CF Gene)
- Child (Does Not Have CF)
- Child (Carrier of CF Gene)
- Child (Carrier of CF Gene)
- Child (Has Cystic Fibrosis)

From i-am-pregnant.com
Many options emerging

Inherited Condition Screening with NGS

Mendelian diseases collectively account for ~20% of infant mortalities and ~18% of pediatric hospitalizations.¹ Currently, molecular tests are available for a little over 25% of these disorders,
Sharing

!=

Worrying
Pre-Womb

<Womb>

Life

Tomb

Post-Tomb
Discovered in 1997, fetal DNA in maternal plasma and serum holds clues to the health of the fetus.

Presence of fetal DNA in maternal plasma and serum

Y M Dennis Lo, Neemi Corbetta, Paul F Chamberlain, Vik Ral, Ian L Sargent, Christopher W G Redman, James S Wainscoat

Summary

Background The potential use of plasma and serum for molecular diagnosis has generated interest. Tumour DNA has been found in the plasma and serum of cancer patients, and molecular analysis has been done on this material. We investigated the equivalent condition in pregnancy—that is, whether fetal DNA is present in maternal plasma and serum.

Methods We used a rapid-boiling method to extract DNA from plasma and serum, DNA from plasma, serum, and nucleated blood cells from 43 pregnant women underwent a sensitive Y-PCR assay to detect circulating male fetal DNA from women bearing male fetuses.

Findings Fetus-derived Y sequences were detected in 24 (80%) of the 30 maternal plasma samples, and in 21 (70%) of the 30 maternal serum samples, from women bearing male fetuses. These results were obtained with only 10 μL of the samples. When DNA from nucleated blood cells extracted from a similar volume of blood was used, only five (17%) of the 30 samples gave a positive Y signal. None of the 13 women bearing female fetuses, and none of the ten non-pregnant control women, had positive results for plasma, serum or nucleated blood cells.

Interpretation Our finding of circulating fetal DNA in maternal plasma may have implications for non-invasive prenatal diagnosis, and for improving our understanding of the fetomaternal relationship.

Lancet 1997; 350: 485–87

Introduction

The passage of nucleated cells between mother and fetus is well recognised. One important clinical application is the use of fetal cells in maternal blood for non-invasive prenatal diagnosis. This approach avoids the risks associated with conventional invasive techniques, such as amniocentesis and chorionic-villus sampling. Substantial advances have been made in the enrichment and isolation of fetal cells for analysis. However, most techniques are time-consuming or require expensive equipment.

There has been much interest in the use of DNA derived from plasma or serum for molecular diagnosis. In particular, there have been reports that tumour DNA can be detected by molecular techniques in the plasma or serum of cancer patients. Such reports prompted us to investigate whether fetal DNA can be detected in maternal plasma and serum.

Methods

Patients

Pregnant women attending the John Radcliffe Hospital (Oxford, UK) were recruited before amniocentesis or delivery. We obtained approval of the project from the Central Oxfordshire Research Ethics Committee. Informed consent was obtained in each case. 5–10 mL maternal peripheral blood was collected into one tube containing edetic acid and one plain tube. For women undergoing amniocentesis, maternal blood was always taken before the procedure, and 10 mL amniotic fluid was also collected for fetal sex determination. For women recruited just before delivery, fetal sex was noted at the time of delivery. Control blood samples were also taken from ten non-pregnant women, and the samples were processed in the same way as those obtained from the pregnant women.
Non-Invasive Prenatal Testing (NIPT)
Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma


Centre for Research into Circulating Fetal Nucleic Acids, Li Ka Shing Institute of Health Sciences, Departments of Chemical Pathology and Obstetrics and Gynaecology, and Centre for Clinical Trials, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; Center for the Study of Biological Complexity and Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284; and Sequenom, Inc., San Diego, CA 92121
plasma, the proportional contribution of plasma DNA sequences from maternal and fetal genomes are evenly represented in maternal DNA fragments in maternal plasma. Furthermore, if both maternal and fetal DNA pools need to be a representative sample of the total DNA pool with maternal plasma, the number of quantifiable sequences would be independent of any particular gene locus to quantify the amount of that chromosome. Therefore, for any fixed volume of maternal plasma, the amount of that chromosome could be more readily achieved.

We previously proposed that the recently available massively parallel genomic sequencing (MPGS) platforms might be used to determine the chromosomal origin of the sequenced DNA from any particular chromosome. In this study, we counted only sequences that could be mapped to just one location in the repeat-masked reference human genome with no mismatch, i.e., sequences that are deemed as a ‘unique’ sequence in the human genome. We termed these sequences as U0–1 sequences for each human chromosome. In this study, we counted only unique sequences generated for the sample, termed % chrN for chromosome N.

The procedural framework of using MPGS for noninvasive prenatal diagnosis of fetal chromosomal aneuploidy is shown in green) than for those with a euploid fetus (cases A–D shown in blue).

% chrN = Unique count for chrN
Total unique count

% chrN z-score for test sample = % chrN_{sample} – mean % chrN_{reference}
S.D. % chrN_{reference}

The procedural framework of using MPGS involves a fetus with the said aneuploidy. We set out to test each pregnancy for noninvasive fetal chromosomal aneuploidy detection in maternal plasma. The z-score for % chr21 was expected when compared with no further fragmentation was required. One plasma DNA molecule was sequenced for 36 bp using the Solexa sequencing-by-synthesis approach. The chromosomal origin of each 36-bp sequence was identified through mapping to the human reference genome by bioinformatics alignment analysis for the Illumina Genome Analyzer, which uses the Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) software. The purpose of the alignment was to simply determine the chromosomal origin of the sequenced DNA molecules in maternal plasma is obtained. In this study, one end of each fragment-ligated DNA fragments both before and after (i.e., two rounds of amplification) a gel electrophoresis-based size fractionation step were not required. The number of sequence reads originating from maternal and fetal plasma DNA (maternal and fetal) molecules were already fragmented and involved in nature, no further fragmentation was required. One plasma DNA (maternal and fetal) molecules were already fragmented and involved in nature (19), no further fragmentation was required.
Noninvasive Whole-Genome Sequencing of a Human Fetus

Jacob O. Kitzman, Matthew W. Snyder, Mario Ventura, Alexandra P. Lewis, Ruolan Qiu, LaVone E. Simmons, Hilary S. Gammill, Craig E. Rubens, Donna A. Santillan, Jeffrey C. Murray, Holly K. Tabor, Michael J. Bamshad, Evan E. Eichler, Jay Shendure

Analysis of cell-free fetal DNA in maternal plasma holds promise for the development of noninvasive prenatal genetic diagnostics. Previous studies have been restricted to detection of fetal trisomies, to specific paternally inherited mutations, or to genotyping common polymorphisms using material obtained invasively, for example, through chorionic villus sampling. Here, we combine genome sequencing of two parents, genome-wide maternal haplotyping, and deep sequencing of maternal plasma DNA to noninvasively determine the genome sequence of a human fetus at 18.5 weeks of gestation. Inheritance was predicted at $2.8 \times 10^6$ parental heterozygous sites with 98.1% accuracy. Furthermore, 39 of 44 de novo point mutations in the fetal genome were detected, albeit with limited specificity. Subsampling these data and analyzing a second family trio by the same approach indicate that parental haplotype blocks of ~300 kilo–base pairs combined with shallow sequencing of maternal plasma DNA is sufficient to substantially determine the inherited complement of a fetal genome. However, ultradeep sequencing of maternal plasma DNA is necessary for the practical detection of fetal de novo mutations genome-wide. Although technical and analytical challenges remain, we anticipate that noninvasive analysis of inherited variation and de novo mutations in fetal genomes will facilitate prenatal diagnosis of both recessive and dominant Mendelian disorders.
Table 1. Summary of sequencing. Individuals sequenced, type of starting material, and final fold coverage of the reference genome after discarding PCR or optical duplicate reads.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Sample</th>
<th>Depth of coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother (I1-M)</td>
<td>Plasma (5 ml, gestational age 18.5 weeks)</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Whole blood (&lt;1 ml)</td>
<td>32</td>
</tr>
<tr>
<td>Father (I1-P)</td>
<td>Saliva</td>
<td>39</td>
</tr>
<tr>
<td>Offspring (I1-C)</td>
<td>Cord blood at delivery</td>
<td>40</td>
</tr>
</tbody>
</table>

A

![Diagram](image1.png)
Using individual sites  Using maternal haplotypes

**B**

Dilution pool whole-genome phasing

**C**

Allele transmitted

Shared paternal-specific

Plasma paternal-specific reads/site

**Figure 1.**

**Experimental approach.** (A) Sequenced individuals in a family trio. (B) Haplotype phasing by whole-genome shotgun. (C) Histogram of plasma paternal-specific reads, compared across Mendelian and non-Mendelian read depth at chromosome Y and alleles of nontransmission (Fig. 1C). In contrast to maternal-only heterozygous reads matching the paternal-specific haplotype, we used a site log-odds test; this amounted to transmitting and, conversely, the lack of such observations as evidence for its transmission and, conversely, the lack of such observations as evidence for nontransmission (Fig. 1C). We anticipated that analogous to the case of maternal transmission, this could be estimated accuracy of 98.7% (Table 2). Although we did not predict paternal transmission at these sites, we anticipate that accuracy could likely be improved assuming 13% fetal content. To assess the observation of one or more shared allele transmitted; green, maternal-specific allele transmitted).
assuming 13% fetal content). To assess these, we performed a site-by-site log-odds test; this amounted to taking the observation of one or more reads matching the paternal-specific IVS1e sequence evidence of its transmission and, conversely, the lack of such observations as evidence of nontransmission (Fig. 1C). In contrast to maternal-only heterozygous sites, this simple site-by-site model was sufficient to correctly predict inheritance at 1.1 × 10^6 paternal-only heterozygous sites with 96.8% accuracy (Table 2). We anticipate that accuracy could likely be improved by deeper sequence coverage of the maternal plasma DNA (fig. S2) or, alternatively, by taking a haplotype-based approach if high-molecular weight genomic DNA from the father is available.

We next considered transmission at sites heterozygous in both parents. We predicted maternal transmission at such shared sites phased using neighboring maternal-only heterozygous sites in the same haplotype block. This yielded predictions at 576,242 of 631,721 (91.2%) of shared heterozygous sites with an estimated accuracy of 98.7% (Table 2).

Although we did not predict paternal transmission at these sites, we anticipate that analogous to the case of maternal transmission, this could be done with high accuracy given paternal haplotypes. We note that shared heterozygous sites primarily correspond to common alleles (fig. S3), which are less likely to contribute to Mendelian disorders in nonconsanguineous populations.

**Fig. 1. Experimental approach.** (A) Sequenced individuals in a family trio. Maternal plasma DNA sequences were ~13% fetal-derived on the basis of read depth at chromosome Y and alleles specific to each parent. WGS, whole-genome shotgun. (B) Inheritance of maternally heterozygous alleles inferred using long haplotype blocks. Among plasma DNA sequences, maternal-specific alleles are more abundant when transmitted (expected, 50% versus 43.5%), but there is substantial overlap between the distributions of allele frequencies when considering sites in isolation (left histogram: yellow, shared allele transmitted; green, maternal-specific allele transmitted). Taking average allele balances across haplotype blocks (right histogram) provides much greater separation, permitting more accurate inference of maternally transmitted alleles. (C) Histogram of fractional read depth among plasma data at paternal-specific heterozygous sites. In the overwhelming majority of cases when the allele specific to the father was not detected, the opposite allele had been transmitted (96.8%, n = 561,552). (D) De novo missense mutation in the gene ACMSD detected in 3 of 93 maternal plasma reads and later validated by PCR and resequencing. The mutation, which is not observed in dbSNP nor among coding exons sequenced from >4000 individuals as part of the National Heart, Lung, and Blood Institute Exome Sequencing Project (http://evs.gs.washington.edu), creates a leucine-to-proline substitution at a site conserved across all aligned mammalian genomes (University of California, Santa Cruz, Genome Browser) in a gene implicated in Parkinson's disease by genome-wide association studies (25).
NGS caveats for clinical sequencing

• Short read technology/aligning to reference genome

• SNV vs indels, CNV……

• Small repeats/repetitive sequence

• Phasing

• Interpretation of unique variants!
Proportion of fetal DNA increases during pregnancy, from

Quantitative Analysis of Fetal DNA in Maternal Plasma and Serum: Implications for Noninvasive Prenatal Diagnosis


Departments of *Chemical Pathology, Obstetrics and Gynecology, and Clinical Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong; and *Department of Hematology, John Radcliffe Hospital, Oxford

Summary

We have developed a real-time quantitative PCR assay to measure the concentration of fetal DNA in maternal plasma and serum. Our results show that fetal DNA is present in high concentrations in maternal plasma, reaching a mean of 25.4 genome equivalents/ml (range 3.3–69.4) in early pregnancy and 292.2 genome equivalents/ml (range 76.9–769) in late pregnancy. These concentrations correspond to 3.4% (range 0.39%–11.9%) and 6.2% (range 2.33%–11.4%) of the total plasma DNA in early and late pregnancy, respectively. Sequential follow-up study of women who conceived by in vitro fertilization shows that fetal DNA can be detected in maternal serum as early as the 7th wk of gestation and that it then increases in concentration as pregnancy progresses. These data suggest that fetal DNA can be readily detected in maternal plasma and serum and may be a valuable source of material for noninvasive prenatal diagnosis.

Cheung et al. 1996). However, most techniques are time consuming, labor intensive, or difficult to implement on a large scale. These limitations prompted us to investigate alternative sources of fetal genetic material for molecular analysis.

There has been much recent interest in the use of DNA derived from plasma or serum, for molecular diagnosis (Boland 1996). In particular, plasma and serum DNA from cancer patients has been shown to contain large quantities of tumor DNA (Chen et al. 1996; Nawroz et al. 1996; Anker et al. 1997). Prompted by such reports, we recently demonstrated that fetal DNA is present in maternal plasma and serum (Lo et al. 1997). Detection of fetal DNA sequences was possible, in 80% and 70% of cases, with just 10 µl of boiled plasma and serum, respectively (Lo et al. 1997).

These observations suggest that maternal plasma/serum DNA may be a useful source of material for the noninvasive prenatal diagnosis of certain genetic disorders. However, a number of important questions need to be answered before clinical application of this new approach can be contemplated. First, fetal DNA in maternal plasma and serum would need to be shown to be...
%cell-free fetal DNA can even reach >30%

29% of cases are between 4% and 8%

4.8% of cases <4% FF

11,225 Commercial Natera NIPT cases

http://www.panoramatest.com/healthcare-provider
With cfDNA, you can know the genome of a fetus well before birth.
However, the %cell-free fetal DNA is possibly dependent on mother’s BMI.
Mosaicism increases with age, which is also a factor in NIPT

DNA Sequencing versus Standard Prenatal Aneuploidy Screening

Diana W. Bianchi, M.D., R. Lamar Parker, M.D., Jeffrey Wentworth, M.D., Rajeevi Madankumar, M.D., Craig Saffer, M.D., Anita F. Das, Ph.D., Joseph A. Craig, M.D., Darya I. Chudova, Ph.D., Patricia L. Devers, M.S., C.G.C., Keith W. Jones, Ph.D., Kelly Oliver, B.S., Richard P. Rava, Ph.D., and Amy J. Sehnert, M.D. for the CARE Study Group


Comments open through March 5, 2014

BACKGROUND
In high-risk pregnant women, noninvasive prenatal testing with the use of massively parallel sequencing of maternal plasma cell-free DNA (cfDNA testing) accurately detects fetal autosomal aneuploidy. Its performance in low-risk women is unclear.

Full Text of Background...

METHODS
At 21 centers in the United States, we collected blood samples from...
Results replicated on different sequencing platform (Proton) as well
## Considerations of Pre-Natal Testing

<table>
<thead>
<tr>
<th></th>
<th>Non-Invasive Pre-Natal Testing (NIPT)</th>
<th>Amniocentesis</th>
<th>Chorionic Villi Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost</strong></td>
<td>$850-$1000</td>
<td>$100-$1500</td>
<td>$399-$599</td>
</tr>
<tr>
<td><strong>Type of Test</strong></td>
<td>Blood draw</td>
<td>Invasive</td>
<td>Invasive</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>&gt;99%</td>
<td>&gt;99%</td>
<td>&gt;98%</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>&gt;99%</td>
<td>&gt;99%</td>
<td>&gt;98%</td>
</tr>
<tr>
<td><strong>TAT (Turn Around Time)</strong></td>
<td>1-2 weeks</td>
<td>2-3 weeks</td>
<td>2-3 weeks</td>
</tr>
<tr>
<td><strong>Earliest Gestational Time for Test</strong></td>
<td>8 weeks</td>
<td>16 weeks</td>
<td>11 weeks</td>
</tr>
<tr>
<td><strong>Trisomy 21 false positives</strong></td>
<td>0.1-0.3%</td>
<td>3.3%</td>
<td>3.6%</td>
</tr>
</tbody>
</table>
Also many options for testing in the womb

<table>
<thead>
<tr>
<th>Company (Test)</th>
<th>Market</th>
<th>Conditions tested</th>
<th>Test price</th>
<th>Tests sold</th>
<th>Market entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ariosa (Harmony)</td>
<td>USA</td>
<td>13, 18, 21, XY aneuploidies</td>
<td>$795</td>
<td>September 2013</td>
<td>May 2012</td>
</tr>
<tr>
<td>BambniTest</td>
<td>China</td>
<td>Trisomies 13, 18, 21</td>
<td>Around $500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGI (NIFTY)</td>
<td>China, Worldwide</td>
<td>Trisomies 13, 18, 21</td>
<td></td>
<td>200,000 until October 2013</td>
<td></td>
</tr>
<tr>
<td>CellScape (Clarity)</td>
<td>USA</td>
<td>In development</td>
<td></td>
<td>None yet</td>
<td></td>
</tr>
<tr>
<td>KellbenX</td>
<td>USA</td>
<td>In development</td>
<td></td>
<td>None yet</td>
<td></td>
</tr>
<tr>
<td>LifeCodexx/ GATC (PraenaTest)</td>
<td>Europe</td>
<td>Trisomies 13, 16, 21</td>
<td>€1,150 ($1,580)</td>
<td>Year to August 2013: 6,000</td>
<td>August 2012</td>
</tr>
<tr>
<td>Natera (Panorama)</td>
<td>USA</td>
<td>13, 18, 21, XY aneuploidies</td>
<td>$2,750</td>
<td></td>
<td>December 2012</td>
</tr>
<tr>
<td>(Materni21 Plus)</td>
<td>USA</td>
<td>16, 21, 22, other</td>
<td>$1,700 for uninsured patients</td>
<td>145,000 until September 2013</td>
<td>October 2011</td>
</tr>
<tr>
<td>Verinata/ Illumina (Verifi)</td>
<td>USA</td>
<td>13, 18, 21, XY aneuploidies</td>
<td>$1,500</td>
<td></td>
<td>March 2012</td>
</tr>
</tbody>
</table>
Pre-Womb

Post-Tomb
cBio portal at MSKCC is a great cancer genomics resource.
Some genes show broad mutational burden
Some show strong tumor-specificity
Well-annotated, browse-able mutations
OncoPrint Profile shows the frequency and overlap of mutations

Case Set: Tumors with sequencing and CNA data: All tumor samples that have CNA and sequencing data (262 samples)

Altered in 211 (81%) of cases

- **IDH1**: 75%
- **IDH2**: 5%
- **TET2**: 1%

Copy number alterations are putative.

Sanger Cancer Gene Census Information:

3 of your query genes are known cancer genes, as cataloged by the Sanger Cancer Gene Census:

<table>
<thead>
<tr>
<th>Gene</th>
<th>Cancer Types (Somatic)</th>
<th>Cancer Types (Germline)</th>
<th>Tissue Types</th>
<th>Mutation Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDH1</td>
<td>glioblastoma</td>
<td></td>
<td>other</td>
<td>missense</td>
</tr>
<tr>
<td>IDH2</td>
<td>GBM</td>
<td>mesenchymal</td>
<td>mesenchymal</td>
<td></td>
</tr>
<tr>
<td>TET2</td>
<td>myelodysplastic syndrome</td>
<td>leukaemia/lymphoma</td>
<td>Mis N, frameshift</td>
<td></td>
</tr>
</tbody>
</table>
Drugs? Easy to find.
Overall Survival Kaplan-Meier Estimate

- **Cases with Alteration(s) in Query Gene(s)**
- **Cases without Alteration(s) in Query Gene(s)**

Logrank Test P-Value: 0.000000
Gene expression level mostly increases linearly with gene copy-number; not always.
Fetal DNA (SRY) is often gone within 1-2 hours of birth

Other cell-free DNA is important for diagnosis
Circulating tumor DNA (ctDNA)

CANCER
Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies

Chetan Bettegowda1,2,*, Mark Sausen1,*,†, Rebecca J. Leary1,*,‡, Isaac Kinde1,*, Yuxuan Wang1,
Nishant Agrawal1,2, Bjarne R. Bartlett1,3, Hao Wang1, Brandon Luber1, Rhoda M. Alani4,
Emmanuel S. Antonarakis1, Nilofer S. Azad1, Alberto Bardelli5,6,7, Henry Brem2, John L. Cameron2,
Clarence C. Lee8, Leslie A. Fecher9,10, Gary L. Gallia2, Peter Gibbs11,12, Dung Le1,3,
Robert L. Giuntoli2, Michael Goggins2, Michael D. Hogarty13, Matthias Holdhoff1,
Seung-Mo Hong2,14, Yuchen Jiao1, Hartmut H. Juhl15, Jenny J. Kim1, Giulia Siravegna16,
Daniel A. Lafer1, Calogero Lauricella16, Michael Lim2, Evan J. Lipson1,
Suely Kazue Nagahashi Marie17, George J. Netto2, Kelly S. Oliner18, Alessandro Olivi2,
Louise Olsson19, Gregory J. Riggins2, Andrea Sartore-Bianchi16, Kerstin Schmidt1, Le-Ming Shih2,
Sueli Mieko Oba-Shinjo17, Salvatore Siena16, Dan Theodorescu20, Jeanne Tie11,
Timothy T. Harkins8, Silvio Veronese16, Tian-Li Wang2, Jon D. Weingart2,
Christopher L. Wolfgang2, Laura D. Wood2, Dongmei Xing2, Ralph H. Huban2, Jian Wu1,21,8,
Peter J. Allen22, C. Max Schmidt23, Michael A. Choti2,8, Victor E. Velculescu1,11,
Kenneth W. Kinzler1,11, Bert Vogelstein1,11, Nickolas Papadopoulos1,11, and Luis A. Diaz Jr1,3,11.
Applications of liquid biopsy

Early detection and monitoring

Brain tumor DNA blocked by blood-brain barrier

Breast cancer
Pancreatic cancer
Colon cancer

Many tumors release DNA fragments that circulate in the bloodstream

Detection of resistance mutations

Targeted therapy

Response to therapy
Selective pressure
Resistant mutation #1
Resistant mutation #2

ctDNA of resistance mutations collected in blood sample

Bettegowda et al., STM, 2014
High variation of ctDNA depending on available vasculature of the tumor

Bettegowda et al., STM, 2014
Higher ctDNA is associated with worse survival
Other cell-free DNA is important to diagnosis during life as well.
Use the donor DNA to discern the proportion of cells circulating.
Successful transplants show donor DNA disappear quickly and stay away.

De Vlaminick et al., STM, 2014
A spike in cell-free donor-derived DNA (cfdDNA) can reveal the risk of an organ rejection.

De Vlaminick et al., STM, 2014

Successful Transplant!

A spike in cell-free donor-derived DNA (cfdDNA) can reveal the risk of an organ rejection.
A clear rejection is always better than a fake promise. ; for people, and organs.

http://www.gurl.com/2013/01/15/rejection-quotes/
Every genome sequenced is a potential piece of preventative medicine
But!
Differences in Accuracy
I Had My DNA Picture Taken, With Varying Results

Kira Peikoff, 28, had her DNA tested by three direct-to-consumer companies, and the results didn’t agree.

By KIRA PEIKOFF
Published: December 30, 2013 | 421 Comments

I Had My DNA Picture Taken, With Varying Results

PSORIASIS

Genes Tested - HLA, IL12B, IL23R, Intergenic_1q21, SPATA2, STAT2, TNFAIP3, TNIP1

Description

This patient has typical genetic risk for psoriasis. This does not mean the patient will or will not develop the disease. This test outcome was determined using genetic laboratory results in conjunction with the patient’s self-reported ethnicity. General preventive measures, such as smoking cessation or stress reduction, could be encouraged.

Elevated Risk

<table>
<thead>
<tr>
<th>NAME</th>
<th>CONFIDENCE</th>
<th>YOUR RISK</th>
<th>AVG. RISK</th>
<th>COMPARED TO AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psoriasis</td>
<td>★★★★</td>
<td>20.2%</td>
<td>10.1%</td>
<td>1.99x</td>
</tr>
</tbody>
</table>

Name of the condition

<table>
<thead>
<tr>
<th>Name of the condition</th>
<th>Your lifetime risk</th>
<th>The normal risk</th>
<th>Your genetic risk level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psoriasis</td>
<td>High genetic risk level</td>
<td>10.1%</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Medium genetic risk level</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kira Peikoff, 28, had her DNA tested by three direct-to-consumer companies, and the results didn’t agree.

By KIRA PEIKOFF
Published: December 30, 2013 | 421 Comments

Genotype Error

- Biological Heterogeneity
- Informatic Noise
- Chemical/Optical Noise

Interpretation Failure
Genome in a Bottle Consortium

National Institute of Standards and Technology (NIST) and the ABRF (Association of Biomolecular Resource Facilities)
Genome in a Bottle (GIAB)
Leveraging NA12878 and other NIST Genomes in a Bottle Standards, like Platinum Genome

http://www.illumina.com/platinumgenomes/
Medically Relevant Variant Calls Still Discordant at High Depth

**WGS Concordance**

<table>
<thead>
<tr>
<th>Region</th>
<th>50x</th>
<th>100x</th>
<th>200x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Silver</td>
<td>85%</td>
<td>90%</td>
<td>94%</td>
</tr>
<tr>
<td>Bronze</td>
<td>54%</td>
<td>68%</td>
<td>82%</td>
</tr>
<tr>
<td>Lead</td>
<td>519</td>
<td>6,589</td>
<td>22,164</td>
</tr>
</tbody>
</table>

**Region % of genome**

<table>
<thead>
<tr>
<th>Region</th>
<th>% of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>95.07%</td>
</tr>
<tr>
<td>Silver</td>
<td>2.95%</td>
</tr>
<tr>
<td>Bronze</td>
<td>0.02%</td>
</tr>
<tr>
<td>Lead</td>
<td>1.96%</td>
</tr>
</tbody>
</table>

**Elliott H. Marquilies**
Each Platform has various sources of noise, and thus Error

- De-Phasing
  - Lagging strand dephasing from incomplete extension
  - Leading strand dephasing from over-extension
- Dark Nucleotides
- Polymerase errors (10^{-5} to 10^{-7})
- Platform-specific errors
  - Illumina more likely to have error after ‘G’
  - PCR-based methods miss GC- and AT-rich regions
Where do these errors come from?
<table>
<thead>
<tr>
<th>Workflow Parameters</th>
<th>Cover Many Steps</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Illumina TruSeq</th>
<th>Illumina Nexterra</th>
<th>LifeTech IonXPress</th>
<th>LifeTech NEBNext</th>
<th>Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purify DNA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum Input</td>
<td>1 μg</td>
<td>5 ng</td>
<td>50 ng</td>
<td>100 ng</td>
<td></td>
</tr>
<tr>
<td>DNA Extraction</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>Non-phenol methods</td>
</tr>
<tr>
<td>Sample QC</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Library Synthesis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reaction Set Up</td>
<td>5 min</td>
<td>5 min</td>
<td>5 min</td>
<td>5 min</td>
<td>Robotics</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>15 min</td>
<td>1 hr</td>
<td>15 min</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End repair reaction</td>
<td>1.5 hr</td>
<td>10 min</td>
<td>15 min</td>
<td>30 min</td>
<td>3'-5' exonuclease &amp; polymerase</td>
</tr>
<tr>
<td>A-tailing reaction</td>
<td>1 hr</td>
<td>10 min</td>
<td>NA</td>
<td>NA</td>
<td>double-A tail</td>
</tr>
<tr>
<td>Adapter ligation</td>
<td>1.5 hr</td>
<td>15 min</td>
<td>15 min</td>
<td></td>
<td>Faster Ligase Enzyme</td>
</tr>
<tr>
<td>Purify Ligation Mixture (Size Selection)</td>
<td>3 hr</td>
<td>10 min</td>
<td>1 hr</td>
<td>1 hr</td>
<td>SPRI beads</td>
</tr>
<tr>
<td>Template Enrichment (PCR)</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
<td>45 min</td>
<td>NA</td>
</tr>
<tr>
<td>Clean up</td>
<td>1 hr</td>
<td>30 min</td>
<td>15 min</td>
<td>15 min</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Total Library Prep.</strong></td>
<td>9.5 hr</td>
<td>1.5 hr</td>
<td>3.5 hr</td>
<td>2.5 hr</td>
<td></td>
</tr>
<tr>
<td><strong>Sequence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library QC and Pooling</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>AATI Analyzer</td>
</tr>
<tr>
<td>Template Annealing (Clustering)</td>
<td>cBot cluster 3h</td>
<td>cBot cluster 3h</td>
<td>IonExpress 3.5 h</td>
<td>IonExpress 3.5 h</td>
<td>NA</td>
</tr>
<tr>
<td>Sequencing</td>
<td>27 hours</td>
<td>27 hours</td>
<td>3 hours</td>
<td>3 hours</td>
<td></td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base Calling</td>
<td>realtime</td>
<td>realtime</td>
<td>realtime</td>
<td>realtime</td>
<td>NA</td>
</tr>
<tr>
<td>Move Data</td>
<td>0-2 days</td>
<td>0-2 days</td>
<td>0-2 days</td>
<td>0-2 days</td>
<td>NA</td>
</tr>
<tr>
<td>Alignment</td>
<td>2-3 days</td>
<td>2-3 days</td>
<td>2-3 days</td>
<td>2-3 days</td>
<td>Isaac/ HPC</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>2-4 hrs</td>
<td>2-4 hrs</td>
<td>2-4 hrs</td>
<td>2-4 hrs</td>
<td>Isaac/ HPC</td>
</tr>
<tr>
<td>SNP annotation</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>1 hr</td>
<td>NA</td>
</tr>
<tr>
<td><strong>~ Shortest Total Time</strong></td>
<td>50 hrs</td>
<td>42 hrs</td>
<td>21 hrs</td>
<td>20 hrs</td>
<td></td>
</tr>
</tbody>
</table>
The Technical Dozen: At least 12 Sources of Variation

(1) DNA/RNA integrity: Sample purity or degradation
(2) Sample complexity: cellular heterogeneity, type of nucleic acid
(3) DNA/cDNA synthesis: random hexamer vs. polyA-primed
(4) Library isolation: Gel excision, column, gel-free
(5) Technical Errors: Machine, Site, Lane, Technician, Library Size
(6) Amplification Cycles or Methods: NuGen, Tn5, Phi29
(7) Input amount: (1, 10, 100, 1000 cells)
(8) Algorithms: for alignment, base-calling, and assembly
(9) Fragment size distribution: Paired-End, Single-End (adaptors)
(10) Ligation Efficiency: Multiplexing/Barcoding and ligase diffs.
(11) Depth of Sequencing: cost/benefit point
(12) RNA/DNA fragmentation: cation, enzymatic, sonication
Fare, T. L. et al.

Effects of atmospheric ozone on microarray data quality.

Many Options for Alignment

- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- ......
Challenges in WGS

"Our big concern is that you can't take the same genome and sequence it and get the same answer twice"

http://cdnwww.genomeweb.com/sequencing/abrf-present-sequencing-platform-performance-study-next-week
<table>
<thead>
<tr>
<th>Data Set</th>
<th>Source</th>
<th>Variants (MM)</th>
<th>Genomes</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Geomes</td>
<td>1000Genomes.org</td>
<td>~25</td>
<td>629</td>
<td>Low coverage, pools</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td><a href="http://www.completegenomics.com/sequence-data/download-data/">www.completegenomics.com/sequence-data/download-data/</a></td>
<td>~19</td>
<td>69</td>
<td>High coverage, individuals</td>
</tr>
</tbody>
</table>
Why so many different mutations?

All data

Complete Genomics
19,061,395

Overlap
15,061,614

1000 Genomes
25,488,488

32 matched genomes

Complete Genomics
15,309,617

Overlap
12,357,915

1000 Genomes
13,029,568
A Comparison of 32 HapMap genomes from CG and 1KGP: Differences of 4-14%

Blue=CG
Red=1KGP

Two simplifying biases of WGS are misleading:

1. We are diploid
2. There is an interesting mutation in a gene
“Genomic Relativity” can create interesting compound heterozygotes

<table>
<thead>
<tr>
<th>chromo</th>
<th>location</th>
<th>Patient1D reads</th>
<th>Patient1R reads</th>
<th>Patient2D reads</th>
<th>Patient2R reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr11</td>
<td>61652197</td>
<td>NA</td>
<td>-G/+GTTCG</td>
<td>NA</td>
<td>-G/-G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paternal</th>
<th>Reference</th>
<th>Maternal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAT1</td>
<td>REF</td>
<td>MAT</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

chr11 61652197: ±GGTCG
ASTONISHED US medics say there's "an overwhelming 99.999 per cent" that a set of 11-month-old Texan twins have been born to different fathers.

ASTONISHED US medics say there's "an overwhelming 99.999 per cent chance" that a set of 11-month-old Texan twins have been born to different fathers.

Babies Justin and Jordan Washington arrived just seven minutes apart, UK tabloid The Sun reports, but were conceived separately because their mother Mia was cheating on partner James Harrison.

A disbelieving Mia, 20, told The Sun: “Out of all the people in America and all the people in the world, this had to happen to me.”

The Texas mum admitted having the affair at the time she conceived, and submitted to a paternity test because the two boys looked so unalike.

The double pregnancy is known technically as “heteropaternal superfecundation”.

It can occur when a woman releases multiple eggs during ovulation, which can be fertilised individually by sperm cells from separate partners.

Genny Thibodeaux, from Clear Diagnostics DNA Lab which did the test, said: “It’s crazy. Most people don’t believe it can happen but it really can.”

The test result means the boys will now be officially referred to as half-brothers.
Indisputable double paternity in dizygous twins

Eloy Girela, M.D., Jose A. Lorente, M.D., J. Carlos Alvarez, M.S., Maria D. Rodrigo, M.D., Miguel Lorente, M.D., Enrique Villanueva, M.D.

Department of Legal Medicine, Faculty of Medicine, University of Granada, Granada, Spain

Received: October 22, 1996; Received in revised form: February 11, 1997; Accepted: February 11, 1997;

DOI: http://dx.doi.org/10.1016/S0015-0282(97)81456-2

Abstract

Objective: To report a case of heteropaternal superfecundation.
Perinatal/Neonatal Case Presentation


Case report: a black and white twin

M J Claas, A Timmermans and H W Bruins

Albinism is an autosomal recessive disorder that is caused by a defective synthesis of melanin, resulting in a generalized reduction of pigmentation in the skin, hair and eyes, and leading to an increased risk of skin cancer and vision problems. We report a case of a 22-year-old primigravida of Negroid origin who delivered dichorial diamniotic twins: two daughters were born with a totally different appearance. The first child had a light brown skin, black curly hair and brown eyes, whereas the second had a striking white skin, red-blond curly hair and blue eyes. Oculocutaneous albinism (OCA) and heteropaternal superfecundation were considered in the differential diagnosis. Genetic testing confirmed the diagnosis of OCA type 2 in the second child. The diagnosis of albinism has clinical implications and must be considered when a black and white twin is born.
Two simplifying biases of WGS are misleading:

1. We are diploid
2. There is an interesting mutation in a gene
3. We have one father
Two simplifying biases of WGS are misleading:

1. We are diploid
2. There is a most-interesting mutation in a gene

On average:
1,129 sites per genome with two non-reference alleles

Additive/Multiplicative/Compensatory Possibilities:
BRCA1 has 2-5 nonsyn SNVs per genome
One AHNAK2 exon has 91 SNVs
6,077 genes with multiple SNVs

Tri-Alleles?
120X more tri-allelic sites found with CG
Genotype Calling - GATK

Alignment
- Sort and index BAM files using SAMtools
- Remove duplicates from BAM files using Picard and re-index
- Re-order BAMs so that contigs are able to run through GATK (using SAMtools ReorderSam)
- Re-align reads around indels with GATK (first run RealignerTargetCreator to find suspicious intervals, then run IndelRealigner)
- Perform base quality score recalibration in GATK (run CountCovariates followed by TableRecalibration).

Variant calling
- Genome-wide raw SNP and indel calls using GATK UnifiedGenotyper
- Filter indel calls using the filters suggested by GATK for high-coverage data (quality, strand bias, depth)
- Filter SNPs calls using the filters suggested by GATK for high-coverage data (quality, strand bias, depth, homopolymer runs, SNP clusters). Also mask SNPs around indels.
- Did not perform variant quality score recalibration because the data quality is high and coverage is sufficient to be able to get good calls using only hard filters.
Genotype calling - Illumina

- **CASAVA** (ELANDv2, multi-seeded aligner):
  - **SNVs**
    - Base calls are ignored where more than 2 mismatches to the reference sequence occur within 20 bases of the call. Note that this filter treats each insertion or deletion as a single mismatch.
    - If the call occurs within the first or last 20 bases of a read then the mismatch limit is applied to the 41 base window at the corresponding end of the read.
    - The mismatch limit is applied to the entire read when the read length is 41 or shorter.
  - **Indels**: callSmallVariants module (local realignment)
- **GROUPER** (large indels, CNVs)
- **Translocations/Duplications**:
  - ClusterMerger (chimeric read pair filter)
  - ReadBroker (builds evidence by merging clusters)
  - AlignContig (contig re-alignment for re-arrangements)
Overlap is High Between the Two Methods

**Patient 2C2**

- **Illumina+GATK**
  - 3,592,127 SNPs
  - 90.6%
- **Illumina-specific**
  - 217,381 SNPs
  - 5.5%
- **GATK-specific**
  - 154,091 SNPs
  - 3.9%

**Patient 2F**

- **Illumina+GATK**
  - 3,592,127 SNPs
  - 90.7%
- **Illumina-specific**
  - 219,568 SNPs
  - 5.5%
- **GATK-specific**
  - 149,147 SNPs
  - 3.8%
## SNP calls comparison

### 2C2

<table>
<thead>
<tr>
<th></th>
<th>Illumina + GATK</th>
<th>Illumina-specific</th>
<th>GATK-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNPs</td>
<td>3,592,127</td>
<td>217,381</td>
<td>154,091</td>
</tr>
<tr>
<td>Ti/Tv</td>
<td>2.11</td>
<td>1.16</td>
<td>1.6</td>
</tr>
<tr>
<td>Novel</td>
<td>5.0%</td>
<td>29.0%</td>
<td>35.0%</td>
</tr>
<tr>
<td>in genes</td>
<td>37.9%</td>
<td>31.9%</td>
<td>24.2%</td>
</tr>
<tr>
<td>in exons</td>
<td>2.1%</td>
<td>1.2%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

### 2F

<table>
<thead>
<tr>
<th></th>
<th>Illumina + GATK</th>
<th>Illumina-specific</th>
<th>GATK-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNPs</td>
<td>3,590,229</td>
<td>219,568</td>
<td>149,147</td>
</tr>
<tr>
<td>Ti/Tv</td>
<td>2.11</td>
<td>1.16</td>
<td>1.61</td>
</tr>
<tr>
<td>Novel</td>
<td>4.9%</td>
<td>29.0%</td>
<td>35.0%</td>
</tr>
<tr>
<td>in genes</td>
<td>38.1%</td>
<td>31.7%</td>
<td>24.8%</td>
</tr>
<tr>
<td>in exons</td>
<td>2.1%</td>
<td>1.8%</td>
<td>2.0%</td>
</tr>
</tbody>
</table>
Some areas are just harder than others

<table>
<thead>
<tr>
<th></th>
<th>Genome</th>
<th>SNVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold</td>
<td>95.07%</td>
<td>94.80%</td>
</tr>
<tr>
<td>silver</td>
<td>2.95%</td>
<td>4.15%</td>
</tr>
<tr>
<td>copper</td>
<td>0.01%</td>
<td>1.05%</td>
</tr>
<tr>
<td>lead</td>
<td>1.96%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
The genome is not stable, but then again, neither is the substrate. The genome reference itself is always in flux!
Remaining Questions for ACMG, AMP, CDC, CAP, NIST

(1) Who does the analysis...and how? BioInformatics
(2) What do you do with Variants of Unknown Significance (VOUSs)
(3) Severe risk discovery – do you report this?
(4) CLIA (Clinical Laboratory Improvement Amendments) and CAP Certification
(5) Reagent cost vs. test cost, ACA
Portable Legal Consent for Common Genomic Research is an Option for people who want to share all of their Omic data

You’ve arrived at the entry point for the Portable Legal Consent for Common Genomic Research (PLC-CGR) study.

From here, you can go ahead and enroll in the study, which involves going through an online “wizard” in which you’ll learn about the risks and benefits of PLC-CGR. Or, if you’re still thinking about this, you can read our Frequently Asked Questions, contact us, read the consent form itself, and in general look around.
A Book with 3 Billion Letters: disease susceptibility, personality, wellness
If a patient dies before information is gathered
VS.
Conclusions

1. Genotype error can be from various sources of heterogeneity:
   a. Biological (mosaicism, somatic mutations, cell-to-cell differences)
   b. Chemical/Optical (reagents, fluorophores, chemistries, enzymes, buffers)
   c. Bioinformatic (Reference genome, aligner, parameters of analysis, variant calling)

2. Heteropaternal superfecundity is possible, albeit rare

3. Sharing of clinical and genomic data can be liberating and helpful, but also potentially risky

4. Nonetheless, longitudinal molecular profiling is the likely future of medicine