

Ryan Tewhey^{1,2}, Jason Warner³, Masakazu Nakano^{1,4}, Brian Libby³, Martina Medkova³, Patricia David³, Steve Kotsopoulos³, Michael Samuels³, J. Brian Hutchison³, Jonathan W. Larson³, Eric J. Topol¹, Michael P. Weiner³, Jeff Olson³, Darren R. Link³, Kelly A. Frazer^{1,4}, Olivier Harismendy^{1,4}.

1. Scripps Genomic Medicine – Scripps Translational Science Institute – The Scripps Research Institute, 3344 N. Torrey Pines Court, La Jolla CA 92037
2. Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093
3. RainDance Technologies, 44 Hartwell Avenue, Lexington, MA 02421
4. Current address: Moores UCSD Cancer Center, University of California, San Diego, La Jolla, CA 92093-0901, USA.

Targeted sequencing of specific loci of the human genome is a promising approach for maximizing the efficiency of second-generation sequencing technologies for population-based studies of genetic variation. We describe a microdroplet PCR, which performs 1.5 million separate amplifications in parallel, as an approach for enriching targeted sequences in the human genome. We initially designed primers to 435 exons of 47 genes that were selected for having a broad spectrum of sequence characteristics. Using this primer set we amplified the same six samples by both microdroplet and traditional singleplex PCR and sequenced the products using the Illumina GAII demonstrating that both methods generate similarly high quality data; 78% of the reads map to targeted sequences, uniform coverage of ~90% of the targeted bases, greater than 99% accuracy in sequence variant calls, and high reproducibility between different samples (r=0.9). We next scaled the microdroplet PCR to 3976 amplicons totaling 1.49 Mb of sequence, sequenced the resulting sample on both the Illumina GAII and Roche 454 platforms, and obtained data with equally high specificity and sensitivity quality. Our results demonstrate that microdroplet technology is well suited for processing DNA for massively parallel amplification of specific subsets of the human genome for targeted sequencing.

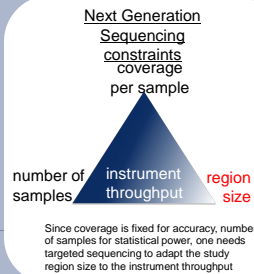
Targeted Sequencing

GOAL: Finding DNA Variants underlying phenotypes and diseases for sequence based association studies

- in candidate regions identified by linkage or Genome-Wide Association Studies
- in candidate genes/pathways for coding variants (exome is 1% of the genome)

SPECIFICATIONS

- **Uniformity** : all targets retrieved in equal amount
- **Sensitivity** : no missing targets
- **Universal** : target can be any piece of the genome
- **Sample to Sample** reproducibility
- **Specificity** : no off-targets
- **Scalability** : high number of samples/targets
- **Accurate** : no allelic representation bias
- **Costs**



Experimental Design

Validation phase : comparison to traditional PCR
457 amplicons (172 kb) sequenced in 6 indexed samples on one lane of Illumina GA

| Samples | 3 caucasian | technical replicate |
|---------|-------------------------------|--|
| | 3 african | influence of unknown DNA variant in PCR |
| Targets | 29 genes in ENCODE | Well annotated region sequence available in 5/6 samples |
| | 8 TRP channel genes | Ability to target related sequences |
| | 11 genes in venous thrombosis | |
| | 435 exons - 457 amplicons | size (119-956 bp) and GC content (33-74%) |

Scale up phase
3976 amplicons (1.49 Mb) sequenced in 1 sample on Illumina GA and Roche 454

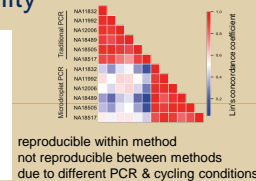
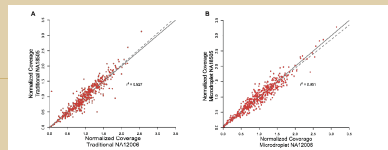
Targeting Specificity

| Sample | PCR Method | Read Number | Bases | Full Amplicon | Missed Bases | Primer Trimmed |
|---------|--------------|-------------|----------|------------------|------------------|------------------|
| NA11832 | Traditional | 149608 | 53.86 Mb | 50.77 Mb (94.3%) | 152.3 kb | 45.54 Mb (84.6%) |
| NA11832 | Microdroplet | 1673982 | 60.26 Mb | 47.50 Mb (78.8%) | 42.55 Mb (70.7%) | |
| NA11992 | Traditional | 1213336 | 43.69 Mb | 40.19 Mb (92.0%) | 36.41 Mb (83.3%) | |
| NA11992 | Microdroplet | 1367394 | 49.23 Mb | 32.42 Mb (65.7%) | 29.40 Mb (59.7%) | |
| NA12006 | Traditional | 1256622 | 45.24 Mb | 42.22 Mb (93.3%) | 38.30 Mb (84.7%) | |
| NA12006 | Microdroplet | 1148454 | 41.34 Mb | 32.04 Mb (77.5%) | 28.94 Mb (70.0%) | |
| NA18905 | Traditional | 1222820 | 44.02 Mb | 40.75 Mb (92.6%) | 37.52 Mb (85.2%) | |
| NA18905 | Microdroplet | 1116948 | 42.21 Mb | 32.58 Mb (77.2%) | 29.93 Mb (71.4%) | |
| NA18517 | Traditional | 836226 | 30.18 Mb | 28.18 Mb (93.4%) | 25.66 Mb (85.0%) | |
| NA18517 | Microdroplet | 587958 | 21.17 Mb | 14.10 Mb (66.6%) | 12.77 Mb (60.3%) | |
| NA18489 | Traditional | 1429868 | 51.48 Mb | 47.44 Mb (92.2%) | 41.91 Mb (81.4%) | |
| NA18489 | Microdroplet | 1985186 | 67.87 Mb | 59.17 Mb (88.1%) | 52.95 Mb (78.0%) | |

- Validation**
- 98% traditional
 - 78% microdroplets
- Scale up**
- 78% Illumina GA
 - 94% Roche 454

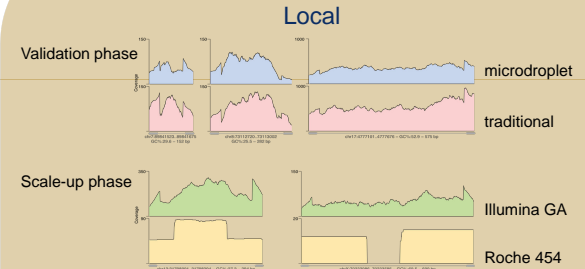
- Non specific mapping reads are scattered around the genome
- More genomic DNA is used in the microdroplet PCR but not carried over in 454 protocol

Reproducibility



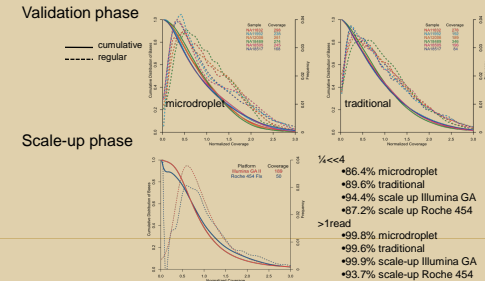
reproducible within method
not reproducible between methods
due to different PCR & cycling conditions

Uniformity of Coverage



Coverage is even along the amplicons.
454 unfragmented strategy creates hills & gaps due to the difference between amplicon size and read length

Global



Combining microdroplet PCR amplified DNA and Illumina GA sequencing to an average depth of 25X that ~ 90% of the bases are covered with 5 or more reads and with ~98% covered by at least one read.

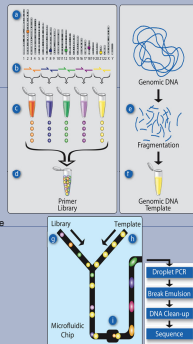
Microdroplet PCR WorkFlow

- Primer Library Generation**
- Identify targeted sequences of interest in the genome.
 - Design and synthesize forward and reverse primer pairs for each targeted sequence
 - Generation of primer pair droplets. A microfluidic chip is used to encapsulate the aqueous PCR primers in inert fluorinated carrier oil with a block-copolymer surfactant to generate the equivalent of a picoliter scale test tube compatible with standard molecular biology.
 - Primer library: primer pair droplets are mixed together so that each library element has an equal representation.

- Genomic DNA Template Mix Preparation**
- Genomic DNA is fragmented into 2 to 4 kb fragments and purified.
 - Purified genomic DNA is mixed together with all of the components of the PCR reaction

- Primer-template merge and PCR**
- Primer Library droplets (~8pL) are dispensed to the microfluidic chip
 - while the Genomic DNA Template is delivered as an aqueous solution and template droplets (~12pL) are formed within the microfluidic chip. The primer pair droplets and template droplets are then paired together in a 1:1 ratio.
 - Paired droplets flow through the channel of the microfluidic chip to pass through a merge area where an electric field induces the two discrete droplets to coalesce into a single PCR droplet (~20 pL). The roughly 1.5 million PCR droplets are collected into a single 0.2 ml PCR tube.

The collection of PCR droplets (PCR Library) is processed in a standard thermal cycler for targeted amplification, followed by breaking the emulsion of PCR droplets to release the PCR amplicons into solution for purification and sequencing.



Accuracy

- Validation Phase** : comparisons to HapMap SNPs (~450 per sample)
- 22/2424 discordant in traditional PCR – 99.09% accurate
 - 22/2390 discordant in microdroplet PCR- 99.08% accurate
 - 50% discordance due to HapMap mistakes (sequencing traces checked)
 - <0.1% (5/2390) due to allelic bias
 - Not sensitive to the presence of unknown variants since :
 - no difference between ENCODE and non-ENCODE region
 - no difference between Caucasian and African

Scale-up phase : comparison to HapMap SNPs

| Sample | Sequencing Platform | HapMap SNPs | Detection Rate | Variant Concordance | Variant SNPs | Discordant SNPs | Common |
|----------|---------------------|-------------|----------------|---------------------|--------------|-----------------|--------|
| NA118858 | Illumina GAII | 2226 | 99.326 | 98.83 | 26 | | |
| NA118858 | 454 Fix | | 92.273 | 98.49 | 31 | 21 | |

extra discordant in 454 due to homopolymer stretches

Conclusion

- **Universal** : all regions are targetable by anchoring Primer in unique sequence
- **Scalable** to high number of targets : tested on 4000 amplicons
- **Scalable** to high number of samples : reproducible and semi-automated
- Few wasted sequenced reads : ~80% **specific**
- **Uniform** Coverage with short reads
- **Accurate** – no allelic bias

Acknowledgments

- X. Wang, K. Post (STSI), O. Iartchouk (Partners HealthCare Center), K. Makowski (Agencourt Bioscience Corporation), N. Schork (STSI), N. Haloz (seqWise)
- US National Institute of Health
 - CTSA grant 1U54RR025204-01
 - Innovative Technologies for Molecular Analysis of Cancer grant 1R21CA125693-01
- Japan Foundation for Aging and Health (MN fellowship)