Improved Exclusion Amplification Chemistry Supports Whole Genome Sequencing of Human TruSeq DNA PCR-free Libraries on Illumina's HiSeq X Ten

Alexandra Kwasniewska*, Peter McInerney, Claire Bevis-Mott, Joe Boutell, Shaun Hunter, Mei Tsan, Martin Fabani, Ark Silbergleit, Michelle Chen, Huihong You, Michael Niziolek, Ernie Guzman, Efstathios Kanterakis, Mikey Pham, Erin Bomati, Molly He, Richard Shen, Susan Tousi and Johanna Whitacre.

1. Abstract

The HiSeq X Ten was designed to unlock the potential of population sequencing, delivering the world's first $1000 human genome at 30x coverage. By using flow cells with billions of patterned nanowells at fixed locations, we have significantly increased the density of clusters that could be achieved by random clustering. To support clustering on patterned flow cells we developed a new cluster chemistry called Exclusion Amplification (ExAmp). Here we describe an improved ExAmp v2 formulation that supports simultaneous seeding and amplification of TruSeq DNA PCR-free libraries, the gold-standard for human whole genome sequencing due to more uniform coverage of regions that are typically underrepresented in amplified libraries (high GC-rich regions, promoters and repetitive regions). This new formulation continues to support amplification of TruSeq DNA Nano PCR libraries and also offers a significant improvement in genomic coverage uniformity. We present data showing improved coverage uniformity for challenging regions of the genome whilst maintaining high quality sequence at genes and exons. Furthermore, by utilizing the Platinum genome data sets from the CEPH pedigree, we have validated our data and found excellent sensitivity and specificity of SNP calling with the improved ExAmp clustering.

2. Product Definition

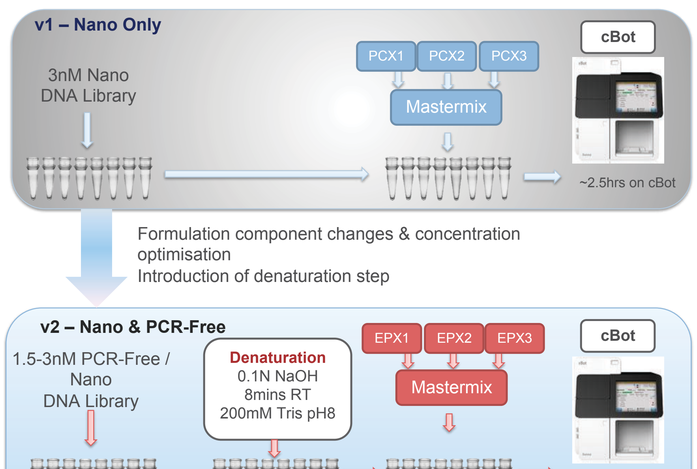
- 1.6 - 1.8T yield generated from 2 x 151+8 run in <3 days, ≥75% Q30+ bases
- Optimized cluster chemistry to improve coverage uniformity at AT/GC extremes
- Support seeding and amplification of TruSeq DNA PCR-free libraries
- Retain compatibility with TruSeq DNA HT Nano (350 bp insert size) libraries
- New RTA/HCS software and cBot recipes

- Overview of HiSeq X kit changes from v1 to the new v2 ExAmp

Assay Component	ExAmp v1	ExAmp v2
Supported Library Preps	• TruSeq Nano350	• TruSeq DNA PCR-free • TruSeq DNA Nano (HT and LT)
ExAmp Formulation	• PCX1, 2, 3 • Reagents support 1 FC	• EPX1,2,3 • Added option for 4 FC configuration
Kit configuration	• 1 kit supports 1 run	Added option for multi-pack (10 runs/kit)
DNA input	• dsDNA, 300pM	• ssDNA workflow - Nano 200-300pM - PCR-free 100-150pM
cBot Protocol	• 2.5 hours	• 3 hours
Supported Indexes	• Single Indexing • TruSeq HT	• Single Indexing • TruSeq LT and HT

3. Workflow changes from v1 to v2 ExAmp clustering

- To support PCR-free clustering DNA must be denatured prior to addition of v2 ExAmp (EPX).



4. High concordance with Platinum Genomes using ExAmp and HiSeq X

•The Platinum Genome Project is primarily an Illumina effort designed to create a comprehensive, genome-wide, gold standard set of variant calls.

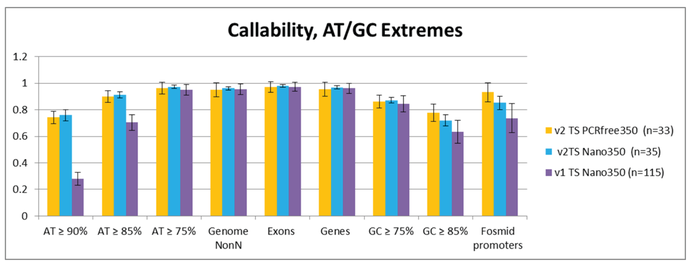
•We used the NA12878 sample from the CEPH Pedigree 1463 for validation of human builds with v2 ExAmp on HiSeq X.

ExAmp / Library type	PG SNP Passing Precision (TP/ TP +FP)	PG SNP Passing Recall (TP/ TP +FN)
v2 TS PCR-free350	0.99559	0.95195
v2 TS Nano350	0.99538	0.95044
v1 TS Nano350	0.99601	0.95051

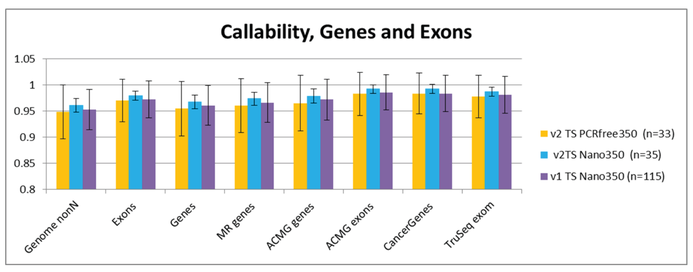
5. v2 ExAmp improves genome coverage on HiSeq X

►Callability (also referred to as coverage) is a number from 0-1 that indicates the proportion of bases in a feature, chromosome or genome that has a PASS position.

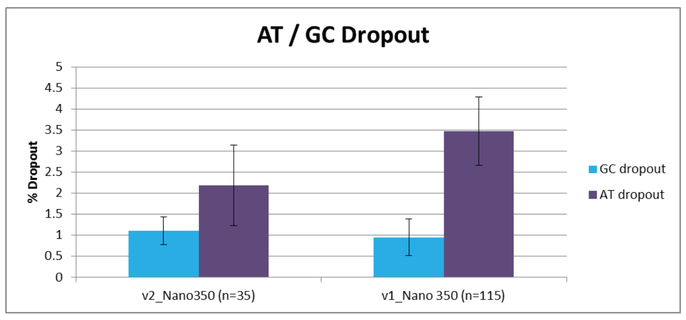
►V2 ExAmp cluster chemistry produces significant improvements in callability of sequences at the extreme AT/GC sequences compared to the V1 formulation.



► Validation of v2 ExAmp chemistry with TruSeq PCR Free Human (350 bp protocol) and TruSeq DNA Nano HT (350bp protocol) found callability to be consistent high for both library types.

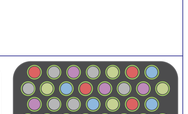


► AT dropout is improved with v2 ExAmp clustering of TruSeq Nano libraries compared to V1 clustering of Nano libraries.



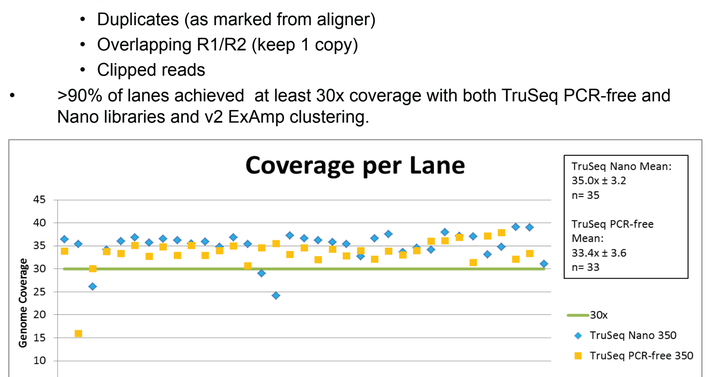
6. %PF differences between random and patterned FC's

- There is a PF penalty for empty wells with patterned FC's, unlike random arrays
  - For patterned FC's % PF is measured against theoretical max density, based on array pitch
- More polyclonal clusters with patterned FC's due to:
  - Lack of filtering due to no template generation
  - Higher DNA input relative to random FC's

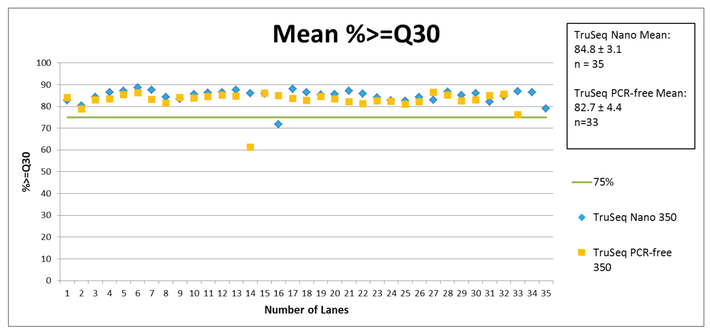
Step	Random clusters	Patterned Clusters
Template generation: Random – spot finding + intensity measurements, low quality clusters filtered out Patterned – no template generation (predefined locations based on array)		
PF determination (grey = non-PF): Random – intensity and chastity measurements over 1st 25 cycles Patterned – intensity and chastity measurements over 1st 25 cycles		

7. Consistent data quality with HiSeq X and v2 ExAmp

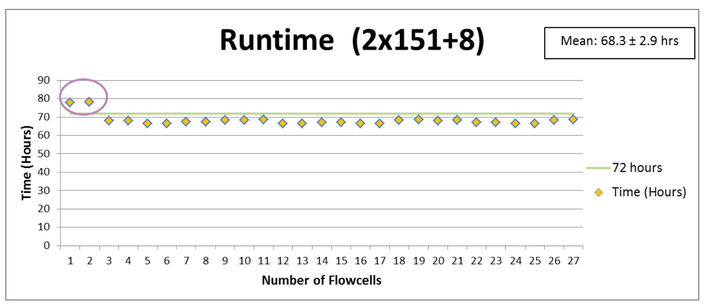
- Target of 112 G should produce 30x coverage per lane.



- Over 95% of TruSeq DNA PCR-free and TruSeq DNA Nano lanes achieved a mean %>=Q30 during the run of >75%.

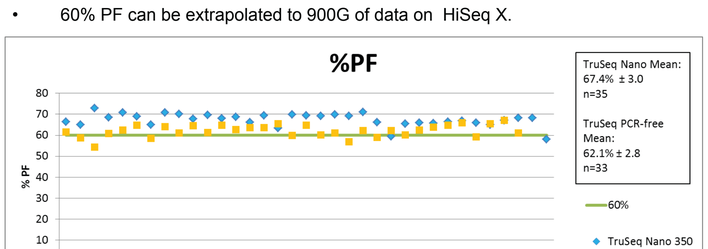


- All but two flowcell's run completed in less than three days (circled), and these took longer as the computer was incorrectly configured, this was easily resolved for subsequent runs.



- 30x Coverage: "*The average coverage over non-N portion of autosomes*"
  - Length: 2,684,578,245 bp
  - Filters:
    - Duplicates (as marked from aligner)
    - Overlapping R1/R2 (keep 1 copy)
    - Clipped reads
- >90% of lanes achieved at least 30x coverage with both TruSeq PCR-free and Nano libraries and v2 ExAmp clustering.

- 60% PF can be extrapolated to 900G of data on HiSeq X.



8. Conclusions

- v2 ExAmp can now support seeding and amplification of TruSeq DNA PCR-free (350bp) libraries in addition to TruSeq DNA Nano (350bp) libraries.
- The new workflow incorporates a NaOH denaturation step and retains the 3 tube ExAmp kit configuration.
- When the same TruSeq DNA Nano library is clustered with either v1 or v2 ExAmp we find that the Platinum genome concordance (recall and precision) is comparable. PCR-free data is also of a similar high quality.
- Significant callability improvements have been made at the AT-rich regions of the genome with the new v2 formulation but the coverage of gene regions is matched across the two versions of ExAmp.
- Quality of runs in maintained with the new chemistry as almost all lanes achieved a mean of >75% %>=Q30 at the full 2x151 + 8 run length in less than 3 days.
- With v2 ExAmp we continue to support a 30x human build per lane with both types of sample prep in over 90% of lanes.
- **v2 ExAmp improves data quality on the HiSeq X Ten, and offers more flexibility to customers with the availability of multi-pack kits and by supporting two TruSeq library kits (Nano and PCR-free).**

illumina®